THE ANNALS *of* MATHEMATICAL STATISTICS

(FOUNDED BY H. C. CARVER)

THE OFFICIAL JOURNAL OF THE INSTITUTE OF MATHEMATICAL STATISTICS

Contents

Vol. XVIII, No. 4 — December, 1947

THE ANNALS OF MATHEMATICAL STATISTICS

The ANNALS OF MATHEMATICAL STATISTICS is published quarterly by the Institute of Mathematical Statistics, Mt. Royal & Guilford Aves., Baltimore 2, Md. Subscriptions, renewals, orders for back numbers and other business communications should be sent to the ANNALS OF MATHEMATICAL STATISTICS, Mt. Royal & Guilford Aves, Baltimore 2, Md., or to the Secretary of the Institute of Mathematical Statistics, P. S. Dwyer, 116 Rackham Hall, University of Michigan, Ann Arbor, Mich.

Changes in mailing address which are to become effective for a given issue should be reported to the Secretary on or before the 15th of the month preceding the month of that issue. The months of issue are March, June, September and December.

Manuscripts for publication in the ANNALS OF MATHEMATICAL STATISTICS should be sent to S. S. Wilks, Fine Hall, Princeton, New Jersey. Manuscripts should be typewritten double-spaced with wide margins, and the original copy should be submitted. Footnotes should be reduced to a minimum and whenever possible replaced by a bibliography at the end of the paper; formulae in footnotes should be avoided. Figures, charts, and diagrams should be drawn on plain white paper or tracing cloth in black India ink twice the size they are to be printed. Authors are requested to keep in mind typographical difficulties of complicated mathematical formulae.

Authors will ordinarily receive only galley proofs. Fifty reprints without covers will be furnished free. Additional reprints and covers furnished at cost.

The subscription price for the ANNALS is $5.00 per year. Single copies $1.50. Back numbers are available at $5.00 per volume or $1.50 per single issue.

COMPOSED AND PRINTED AT THE
WAVERLY PRESS, INC.
BALTIMORE, MD., U. S. A.

Entered as second-class matter at the Post Office at Baltimore, Maryland, under the act of March 3, 1879

ON OPTIMUM TESTS OF COMPOSITE HYPOTHESES WITH ONE CONSTRAINT[1]

By E. L. Lehmann

University of California, Berkeley

Summary. This paper is concerned with optimum tests of certain composite hypotheses. In section 2 various aspects of a theorem of Scheffé concerning type B_1 tests are discussed. It is pointed out that the theorem can be extended to cover uniformly most powerful tests against a one-sided set of alternatives. It is also shown that the method for determining explicitly the optimum test region may in certain cases be reduced to a simple formal procedure. These results are used in section 3 to obtain optimum tests for the composite hypothesis specifying the value of the circular serial correlation coefficient in a normal distribution. A surprising feature of this example is the fact that for the simple hypothesis obtained by specifying values for the nuisance parameters no test with the corresponding optimum properties exists.

In section 4 the totality of similar regions is obtained for a large class of probability laws which admit a sufficient statistic. Some composite hypotheses concerning exponential and rectangular distributions are treated in section 5. It is proved that the likelihood ratio tests of these hypotheses have various optimum properties.

1. Introduction. In developing tests for a class of hypotheses three phases may be distinguished. First, tests are obtained which are intuitively appealing; next, it is shown that these tests have certain attractive features; finally, it is proved that they are "best possible" tests.

In dealing with parametric hypotheses, the likelihood ratio principle is frequently used to obtain a reasonable test. For many of the tests so derived for normal and exponential distributions, the question of bias has been investigated. In most cases unbiasedness has been established; in the other cases, usually a test based on the same criterion but with the boundaries shifted, can be proved to be unbiased. Other desirable properties which likelihood ratio tests have been shown to possess, relate to the asymptotic behaviour of these tests as the sample sizes tend to infinity. An interesting problem which does not seem to have been treated is the question of admissibility of likelihood ratio tests, a test being admissible if its power can not be improved upon uniformly by any other test of the same level of significance.

Investigations of optimum tests of composite hypotheses have been carried through for many hypotheses concerning normal distributions. When the hypothesis specifies the value of one parameter (hypothesis with one constraint), uniformly most powerful one-sided and type B_1 (uniformly most powerful un-

[1] Presented at a meeting of the Institute of Mathematical Statistics in San Diego, June, 1947.

biased) tests have been obtained. When the number of constraints is larger than one, not so much can be expected. It has been shown for some of the tests in this class that they have maximum average power uniformly over a family of surfaces in the parameter space, or that they are uniformly most powerful with respect to the subclass of tests whose power depends only on some function of the parameters. (All optimum properties mentioned are relative only to the class of all similar regions. This will be so throughout the paper and will usually not be stated explicitly).

Two methods for finding uniformly most powerful or uniformly most powerful one-sided regions and type B_1 tests, if they exist are known. Neyman and Pearson [1] developed a method for determining all similar regions, and applied it to obtain uniformly most powerful one-sided tests of certain hypotheses. Neyman [2, 3] extended the method to obtain, for certain hypotheses, the class of all bisimilar (unbiased similar) regions, and Scheffé [4], developing the method further, proved the existence of type B_1 tests for an important class of hypotheses.

A different method for obtaining all similar and bisimilar regions was devised by P. L. Hsu and was used by him and other writers to prove various optimum properties of the likelihood ratio tests for the general linear hypothesis, of Hotelling's T^2 and of other tests [5, 6, 7, 8].

In the present paper we are concerned with applications of these two methods to composite hypotheses with one constraint. However, the applicability is not so restricted. In fact, the second method has been used mainly in connection with composite hypotheses with many constraints, and the author believes it to be suitable also for deriving optimum classification procedures. An essential restriction of both methods seems to be that a set of sufficient statistics must exist with respect to the parameters involved: with respect to the nuisance parameters so that all similar regions can be found, with respect to the parameters specified by the hypothesis so that there exists a best of all similar regions.

Extensions of the existing theory based on the first method are obtained in section 2, and the theory is applied in section 3 to a hypothesis concerning a multivariate normal distribution. Sections 4 and 5 are concerned with applications of the second method to problems to most of which the earlier method is not applicable, in particular to hypotheses concerning exponential and rectangular distributions, hitherto only treated from the likelihood ratio point of view.

2. On the theory of optimum tests.

2.1 *One-sided tests.* In an interesting paper [4], Scheffé determined the type B and type B_1 tests of a certain class of composite hypotheses specifying the value θ_0 of a parameter θ in the presence of nuisance parameters.

Scheffé's results can, in an obvious way, be extended to cover one-sided sets of alternatives. To show this, consider the method used in [4]. Under certain assumptions all tests[2] are found which satisfy the two conditions:

[2] The terms "the test w" and "the region [of rejection] w" will be used interchangeably.

(a) The power function β_w at θ_0 has a preassigned value ϵ (the level of significance), independent of the nuisance parameters;

(b) the power function at θ_0 has derivative 0. (Condition of unbiasedness).

Then that test w_0 is determined for which, of all those satisfying (a) and (b),

(c) the second derivative at θ_0, $\beta''_w(\theta_0)$, is as large as possible.

By definition w_0 is a type B test. Under a certain additional assumption (this is the convexity assumption $\frac{\partial^2 g}{\partial y_1^2} > 0$ of Scheffé's Theorem 2) it is shown that of all tests satisfying (a) and (b), w_0 has maximum power against all alternatives, i.e. is of type B_1.

If now we want to maximize the power against only the one-sided set of alternatives, $\theta > \theta_0$, we determine that test w_1 of all those satisfying (a), for which

(d) the first derivative at θ_0, $\beta'_w(\theta_0)$, is as large as possible.

Under a certain additional assumption (in Scheffé's notation this would be the monotonicity assumption $\frac{\partial g}{\partial y_1} > 0$) it can then be shown that of all tests satisfying (a), w_1 has maximum power against all alternatives $\theta > \theta_0$, (it also has minimum power against all alternatives $\theta < \theta_0$), i.e. w_1 is uniformly most powerful against alternatives $\theta > \theta_0$. We shall not carry through the discussion in detail since Scheffé's argument applies step by step, with only the obvious changes.

2.2 *Determination of the boundaries.* Let $X_1, \cdots, X_n$ be n random variables with a joint probability density function p, depending on parameters θ_1 and $\theta = (\theta_2, \cdots, \theta_l)$. We shall denote the probability density function of a set of random variables $X_1, \cdots, X_n$ whose distribution depends on a parameter θ by $p(x_1, \cdots, x_n \mid \theta)$ or simply by $p(x_1, \cdots, x_n)$ when the dependence on θ is clear from the context. The set of points $(x_1, \cdots, x_n)$ for which

$$p(x_1, \cdots, x_n \mid \theta)$$

is positive we shall denote by $W_+(\theta)$.

Let

$$\varphi_i(x_1, \cdots, x_n) = \frac{\partial}{\partial \theta_i} \log p(x_1, \cdots, x_n \mid \theta_1, \theta)\,|_{\theta_1 = \theta_1^0}, \qquad (i = 1, \cdots, l), \tag{2.1}$$

and let the random variable Φ_i be defined by

$$\Phi_i = \varphi_i(X_1, \cdots, X_n). \tag{2.2}$$

Then for testing the hypothesis H: $\theta_1 = \theta_1^0$, under the assumptions stated by Scheffé, the type B_1 test w_0 is defined by the inequalities

$$\varphi_1 < k_1, > k_2 \qquad (k_1 < k_2) \tag{2.3}$$

where k_1, k_2 depend on θ_1^0, θ, $\varphi_2, \cdots, \varphi_l$ and are determined by the two equations[3]

$$\int_{k_1}^{k_2} \varphi_1^s p(\varphi_1, \cdots, \varphi_l)\, d\varphi_1 = (1 - \epsilon) \int_{-\infty}^{\infty} \text{ same} \qquad (s = 0, 1). \tag{2.4}$$

[3] Although k_1 and k_2 may depend on θ, w_0 is independent of θ, as was shown in [4].

The equations (2.3) and (2.4) are not suitable for the determination of the boundary of w_0. The variables have to be transformed so as to obtain for w_0 an expression from which the calculation of the boundaries becomes feasible, (cf. [9]). This part of the work may be formalized in the following theorem.

THEOREM 1. *Let*

$$U = f(\Phi_1, \Phi_2, \cdots, \Phi_l)$$

$$V_i = g_i(\Phi_2, \cdots, \Phi_l), \qquad (i = 2, \cdots, l), \tag{2.5}$$

be a system of functions, continuously differentiable and with non-vanishing Jacobian almost everywhere, and such that

(i) *U is a linear function of Φ_1*

$$U = a\Phi_1 + b \tag{2.6}$$

with coefficients which may depend on $\Phi_2, \cdots, \Phi_l$ and such that[4] *$a(\Phi_2, \cdots, \Phi_l) > 0$,*

(ii) *it is possible to solve for $\Phi_2, \cdots, \Phi_l$ in terms of the V's,*

(iii) *under the hypothesis H, U is distributed independently of*

$$V = (V_2, \cdots, V_l).$$

Then the region w_0 is equivalent to the region

$$u < c_1, > c_2 \qquad (c_1 < c_2) \tag{2.7}$$

where c_1, c_2 are determined by

$$\int_{c_1}^{c_2} u^s p(u)\, du = (1 - \epsilon) \int_{-\infty}^{\infty} u^s p(u)\, du \qquad (s = 0, 1). \tag{2.8}$$

PROOF.

$$\begin{aligned} p(\varphi_1, \varphi_2, \cdots, \varphi_l) &= p(u, v_2, \cdots, v_l) \cdot \left| \frac{\partial(u, v_2, \cdots v_l)}{\partial(\varphi_1, \cdots, \varphi_l)} \right| \\ &= p(u) \cdot p(v_2, \cdots, v_l) \frac{\partial u}{\partial \varphi_1} \cdot \left| \frac{\partial(v_2, \cdots, v_l)}{\partial(\varphi_2, \cdots, \varphi_l)} \right|. \end{aligned} \tag{2.9}$$

But

$$\begin{aligned} u &= a(\varphi_2, \cdots, \varphi_l) \cdot \varphi_1 + b(\varphi_2, \cdots, \varphi_l) \\ &= \alpha(v_2, \cdots, v_l) \cdot \varphi_1 + \beta(v_2, \cdots, v_l) \end{aligned} \tag{2.10}$$

so that (2.4) reduces to

$$\begin{aligned} \int_{c_1(v_2, \cdots, v_l)}^{c_2(v_2, \cdots, v_l)} \left(\frac{u - \beta}{\alpha} \right)^s p(u) p(v_2, \cdots, v_l)\, du \\ = (1 - \epsilon) \int_{-\infty}^{\infty} \text{same} \qquad (s = 0, 1) \end{aligned} \tag{2.11}$$

[4] A similar theorem holds when we assume $a(\Phi_2, \ldots, \Phi_l) < 0$.

and hence to

$$\int_{c_1(v_2,\cdots,v_l)}^{c_2(v_2,\cdots,v_l)} u^s p(u)\,du = (1-\epsilon)\int_{-\infty}^{\infty} \text{same} \qquad (s=0,1) \tag{2.12}$$

which shows c_1 and c_2 to be independent of the v's. Also obviously (2.3) transforms into (2.7) which completes the proof.

If U is such that its distribution (when $\theta_1 = \theta_1^0$) is independent of θ, c_1 and c_2 of theorem 1 will depend only on the data of the problem: ϵ, n, θ_1^0. However, the existence of constants c_1 and c_2 satisfying (2.8) still has to be proved. We may show more generally the existence of k_1 and k_2 satisfying (2.4). A proof is immediately supplied by an argument which was used by Neyman [10] and Wald [11] to prove the existence of type A tests, and which may be stated in the following

LEMMA. *Let* $0 < \alpha < 1$, *let* $f(x) \geq 0$ *and* $\int_{-\infty}^{\infty} x^s f(x)\,dx < \infty$ *for* $s = 0, 1$. *Then there exist* A, B *such that*

$$\int_A^B x^s f(x)\,dx = \alpha \int_{-\infty}^{\infty} x^s f(x)\,dx \qquad (s=0,1). \tag{2.13}$$

3. Testing for circular serial correlation in a normal population. We now apply the results of the previous section to obtain the optimum tests (i.e. uniformly most powerful against the one-sided set of alternatives, type B_1 in the two-sided case) for the hypothesis specifying the value of the circular serial correlation coefficient in the normal population considered by Dixon [12]. (For the literature on testing for non-circular serial correlation in normal populations cf. [12]).

We assume

$$p(x_1, \cdots, x_n) = \frac{1-\delta^n}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i-\xi)-\delta(x_{i+1}-\xi)]^2\right] \tag{3.1}$$

where $x_{n+1} = x_1$ and $|\delta| < 1$, and we test the hypothesis $\delta = \delta_0$. For testing purposes only the value $\delta_0 = 0$ is of interest presumably, however, the family of tests for arbitrary δ_0 is required for estimating δ by means of confidence intervals, and therefore the more general hypothesis is considered.

Making a transformation in one of the parameters we write

$$p(x_1, \cdots, x_n) = C(\delta, \alpha)\exp\left[\alpha\left[(1+\delta^2)\sum_{i=1}^{n}(x_i-\xi)^2 - 2\delta\sum_{i=1}^{n}(x_i-\xi)(x_{i+1}-\xi)\right]\right] \tag{3.2}$$

where in the notation of the previous section $\theta_1 = \delta$, $\theta_2 = \alpha$, $\theta_3 = \xi$.

THEOREM 2. *For testing the hypothesis* $\delta = \delta_0$ *for the distribution* (3.2)
(a) *the type* B_1 *test exists and is given by*

$$r < r_1, > r_2 \tag{3.3}$$

where

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.4}$$

and where r_1 and r_2 are determined by

$$\int_{r_1}^{r_2}\left(\frac{r}{1 + \delta_0^2 - 2\delta_0 r}\right)^s p(r)\,dr = (1 - \epsilon)\int_{-\infty}^{\infty} \text{same} \qquad (s = 0, 1). \tag{3.5}$$

(b) *the uniformly most powerful similar region for testing H against the alternatives $\delta > \delta_0$ exists and is given by*

$$r > r' \tag{3.6}$$

where r' is determined by

$$\int_{-\infty}^{r'} p(r)\,dr = (1 - \epsilon)\int_{-\infty}^{\infty} p(r)\,dr.^5 \tag{3.7}$$

PROOF. We compute

$$\begin{aligned}
\varphi_1 &= C_1(\delta_0, \alpha) + 2\alpha[\delta_0\Sigma(x_i - \xi)^2 - \Sigma(x_i - \xi)(x_{i+1} - \xi)] \\
\varphi_2 &= C_2(\delta_0, \alpha) + (1 + \delta_0^2)\Sigma(x_i - \xi)^2 - 2\delta_0\Sigma(x_i - \xi)(x_{i+1} - \xi) \\
\varphi_3 &= -2n\alpha(1 - \delta_0^2)(\bar{x} - \xi).
\end{aligned} \tag{3.8}$$

There is no difficulty in checking the conditions of Scheffé's theorems [4].

Next we apply Theorem 1 of the previous section, and define

$$\begin{aligned}
V_2 &= (1 + \delta_0^2)\Sigma(X_i - \bar{X})^2 - 2\delta_0\Sigma(X_i - \bar{X})(X_{i+1} - \bar{X}) \\
V_3 &= \bar{X} - \xi \\
U &= \frac{\Sigma(X_i - \bar{X})(X_{i+1} - \bar{X})}{V_2}.
\end{aligned} \tag{3.9}$$

Conditions (i) and (ii) of Theorem 1 are easily seen to be satisfied. To show that U is independent of $V = (V_2, V_3)$ we employ arguments which have recently been used by various authors in a number of similar problems (cf. [13, 14, 15]).

It is seen that an orthonormal transformation exists:

$$X_1, \cdots, X_n \to Y_1, \cdots, Y_n$$

such that

$$\begin{aligned}
\sqrt{n}\,\bar{X} &= Y_1 \\
\sum_{i=1}^{n}(X_i - \bar{X})(X_{i+1} - \bar{X}) &= \sum_{i=2}^{n}\lambda_i Y_i^2 \\
\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=2}^{n} Y_i^2.
\end{aligned} \tag{3.10}$$

---

[5] A corresponding result holds for the other one-sided case.

Under H the Y's are distributed with probability density

$$(3.11)\quad p(y_1, \cdots, y_n) = C(\delta_0, \alpha) \exp\left[\alpha\left[k(y_1 - \sqrt{n}\,\xi)^2 + \sum_{i=2}^{n} \mu_i y_i^2\right]\right]$$

where $k, \mu_2, \cdots, \mu_n$ depend on δ_0 and where the μ's are all positive. Introducing new variables

$$(3.12)\qquad Z_i = \sqrt{\mu_i}\, Y_i, \qquad (i = 2, \cdots, n),$$

and, then, generalized polar coordinates in the space of the Z's,

$$(3.13)\qquad R = \sqrt{\sum_{i=2}^{n} Z_i^2}, \qquad \Psi_1, \cdots, \Psi_{n-2}$$

we see that Y_1, R and $\Psi_1, \cdots, \Psi_{n-2}$ are completely independent. Also

$$V_2 = R^2, \qquad V_3 = \frac{1}{\sqrt{n}}(Y_1 - \xi)$$

while U, being homogeneous of degree 0 in the Z's, is a function of the Ψ's only. This proves that U, V_2 and V_3 are completely independent. The type B_1 test of H is therefore given by

$$(3.14)\quad u = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_{i+1} - \bar{x})}{(1 + \delta_0^2)\sum_{i=1}^{n} (x_i - \bar{x})^2 - 2\delta_0 \sum_{i=1}^{n} (x_i - \bar{x})(x_{i+1} - \bar{x})} < c_1, > c_2$$

where c_1 and c_2 are determined by

$$(3.15)\qquad \int_{c_1}^{c_2} u^s p(u)\, du = (1 - \epsilon)\int_{-\infty}^{\infty} u^s pu()\, du \qquad (s = 0, 1).$$

We still have to show that this test is equivalent to the one defined by (3.3) and (3.5). For $\delta_0 = 0$ this is trivial. Let us assume $\delta_0 < 0$. (The other case goes through similarly.) The inequality $u < c_1$ is equivalent to

$$(3.16)\qquad (1 + 2\delta_0 c_1)\Sigma(x_i - \bar{x})(x_{i+1} - \bar{x}) < (1 + \delta_0^2)\Sigma(x_i - \bar{x})^2$$

and hence to

$$(3.17)\qquad \frac{\Sigma(x_i - \bar{x})(x_{i+1} - \bar{x})}{\Sigma(x_i - \bar{x})^2} < w_1$$

provided $1 + 2c_1\delta_0 > 0$. Suppose $1 + 2c_1\delta_0 \leq 0$, i.e. $c_1 \geq -\frac{1}{2\delta_0}$. Then[6]

$$(3.18)\quad P\{U < c_1\} \geq P\left\{U < -\frac{1}{2\delta_0}\right\} = P\{0 < \Sigma(X_i - \bar{X})^2\} = 1$$

[6] We denote the probability of an event A by $P\{A\}$.

i.e. $P\{U < c_1\} = 1$ which would contradict (3.15). Similarly if $1 + 2c_2\delta_0 \leq 0$ we would have $P\{U > c_2\} = 0$ and hence our test would be one-sided and therefore not unbiased. The inequalities $u < c_1$, $> c_2$ are thus equivalent to the inequalities $r < r_1$, $> r_2$ and since

$$u = \frac{r}{1 + \delta_0^2 - 2\delta_0 r},$$

(3.5) also follows.

The existence of type B_1 and uniformly most powerful one-sided tests of the hypothesis H is rather surprising. For when α and ξ are assumed known, neither the type A_1 test nor the uniformly most powerful one-sided test of the simple hypothesis H': $\delta = \delta_0$ exists. This is easily seen by determining the most powerful and the most powerful unbiased test against a specific alternative δ_1 for the hypothesis H' in the population

$$p(x_1, \cdots, x_n) = \frac{1 - \delta^n}{(\sqrt{2\pi})^n} \exp\left[-\tfrac{1}{2}[(1 + \delta^2)\Sigma x_i^2 - 2\delta\Sigma x_i x_{i+1}]\right]. \tag{3.19}$$

The distribution of the criterion R was obtained by R. L. Anderson [16] (see also [17]) for the case $\delta = 0$. Madow [15] using Anderson's result found the distribution for arbitrary δ. (Approximations to the distribution have been studied by various authors; for the literature on this cf. [18]. Recently Hsu [19] obtained an asymptotic expansion.) A direct derivation for arbitrary δ may be based on the following theorem of Cramér, which was communicated to the author by Dr. P. L. Hsu.

THEOREM 3. *(Cramér)*[7]. *If X, Y are two random variables, (not necessarily independent), $Y > 0$, then*

$$P\left\{\frac{X}{Y} \leq x\right\} = -\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\varphi_x(t) - \psi(t)}{it}\, dt \tag{3.20}$$

where φ_x and ψ are the characteristic functions of $X - xY$ and Y respectively, provided

$$\int_{-\infty}^{\infty} \left|\frac{\varphi_x(t) - \psi(t)}{t}\right| dt < \infty. \tag{3.21}$$

THEOREM 4. *If*

$$p(x_1, \cdots, x_n) = \frac{1 - \delta^n}{(\sqrt{2\pi}\,\sigma)^n} \cdot \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i - \xi) - \delta(x_{i+1} - \xi)]^2\right], \qquad (x_{n+1} = x_1) \tag{3.22}$$

[7] Differentiated forms of the theorem were given by R. C. Geary [*Jour. Roy. Stat. Soc.* Vol. 107 (1944) p. 56] and H. Cramér [Exercise 6 on p. 317 of *Mathematical Methods of Statistics*. Princeton Univ. Press (1946)].

and if

$$R = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(X_{i+1} - \bar{X})}{\sum_{i=1}^{n} (X_i - X^2)}, \tag{3.23}$$

then

$$P\{R > r\} = \frac{2^{n+1/2}}{n} \frac{1 - \delta^n}{(1 - \delta)(1 + \delta^2 - 2\delta r)} \cdot \sum_j \frac{(-1)^{j+1}(\lambda_j - r)^{n-3/2} \sin \frac{j\pi}{n} \sin \frac{2j\pi}{n}}{1 + \delta^2 - 2\delta\lambda_j} \tag{3.24}$$

where the summation is extended over all integer j, $1 \leq j \leq \frac{n}{2}$, *for which* $\lambda_j > r$, *and where*

$$\lambda_j = 2 \cos \frac{2j\pi}{n}. \tag{3.25}$$

The proof of this theorem from Theorem 3 is straightforward and only will be indicated here. If X and Y denote the numerator and denominator of R respectively, the characteristic functions of Y and $X - rY$ may be obtained by the method of circulants (cf. [12, 17]). The integral on the right hand side of (3.20) is then easily evaluated by the theory of residues when n is odd. In the case that n is even, the integrand has two branchpoints, one in the lower and one in the upper half plane. These may be separated, and then again the method of residues may be applied.

4. Similar regions. The problem of finding all regions similar to the sample space with respect to a parameter θ was solved by Neyman and Pearson [1] for a certain class of probability laws. In a later paper Neyman proved ([20] proposition IX) that if there exists a sufficient statistic T for a parameter θ, then w is similar with respect to θ if it has the following structure: For the intersection $w(t)$ of w with the surface $T = t$, the relative probability of $w(t)$ given $T = t$ has a constant value independent of t. We shall show in this section that for a large class of probability laws which admit a sufficient statistic for θ the regions with the above structure are the only ones that are similar with respect to θ.

We consider samples from a univariate distribution and we distinguish three cases as one, both or neither of the extremes of the range of the distribution depend on the parameter θ. For the first of these cases (cf. Pitman [21]) we consider samples from a distribution with probability density

$$p(x) = \frac{f(x)}{g(\theta)}, \qquad k(\theta) \leq x \leq c, \tag{4.1}$$

where $k(\theta)$ is a strictly monotone continuous function of θ and where c may be infinite. Introducing a new parameter $\delta = k(\theta)$ the distribution of a sample from (4.1) is given by

$$p(x_1, \cdots, x_n) = \frac{f(x_1) \cdots f(x_n)}{b(\delta)}, \qquad \delta \leq x_i \leq c. \tag{4.2}$$

To obtain the totality of regions w similar with respect to δ let us denote by $W_1, \cdots, W_n$ the portions of the sample space where the smallest of the x's is $x_1, \cdots, x_n$ respectively. For any region w denote by w_k' the intersection of w with W_k. Consider a transformation carrying $W_2, \cdots, W_n$ into W_1, letting $y_1 = \min(x_1, \cdots, x_n)$ and letting in W_k:

$$y_2 = x_1, y_3 = x_2, \cdots, y_k = x_{k-1}, y_{k+1} = x_{k+1}, \cdots, y_n = x_n. \tag{4.3}$$

Denote by w_k the image of w_k' under this transformation. The condition that w be similar with respect to δ,

$$\int_w \frac{f(x_1) \cdots f(x_n)}{b(\delta)}\, dx_1 \cdots dx_n \overset{(\delta)}{\equiv} \epsilon, \tag{4.4}$$

may be written in the form

$$\begin{aligned} &\int_\delta^c \frac{f(y_1)}{b(\delta)} \left\{ \sum_{k=1}^n \int_{w_k(y_1)} f(y_2) \cdots f(y_n)\, dy_2 \cdots dy_n \right\} dy_1 \\ &\qquad \overset{(\delta)}{\equiv} n\epsilon \int_\delta^c \frac{f(y_1)}{b(\delta)} \left\{ \int_{(W(y_1))} f(y_2) \cdots f(y_n)\, dy_2 \cdots dy_n \right\} dy_1 \end{aligned} \tag{4.5}$$

where $W(y_1)$ denotes the region $y_1 \leq y_i \leq c$, $(i = 2, \cdots, n)$, that is, the region of variation of $y_2, \cdots, y_n$ given y_1, and where $w_k(y_1)$ denotes the region of variation of $y_2, \cdots, y_n$ given y_1 and w_k. From (4.5) we obtain

$$\frac{1}{b(\delta)} \int_\delta^c f(y_1)\psi(y_1)\, dy_1 \overset{(\delta)}{\equiv} 0 \tag{4.6}$$

where

$$\begin{aligned} \psi(y_1) = &\sum_{k=1}^n \int_{w_k(y_1)} f(y_2) \cdots f(y_n)\, dy_2 \cdots dy_n \\ &- n\epsilon \int_{y_1}^c \cdots \int_{y_1}^c f(y_2) \cdots f(y_n)\, dy_2 \cdots dy_n. \end{aligned} \tag{4.7}$$

But (4.6) implies

$$\psi(y_1) = 0 \text{ almost everywhere} \tag{4.8}$$

and since we can only determine w up to a set of measure 0, we may omit the qualification in (4.8). Therefore a necessary and sufficient condition for w to be similar is

$$(4.9)\qquad \frac{1}{n\left[\int_{y_1}^{c} f(y)\,dy\right]^{n-1}} \sum_{k=1}^{n} \int_{w_k(y_1)} f(y_2)\cdots f(y_n)\,dy_2\cdots dy_n = \epsilon$$

for all y_1.

To see more clearly the structure of these regions, let us take $n = 2$. Equation (4.9) states that on each of the broken lines of Fig. 1 the relative probability of $w = w_1' + w_2'$ given $Y_1 = y_1$ is ϵ, where the decomposition of this probability into its two components may vary with y_1.

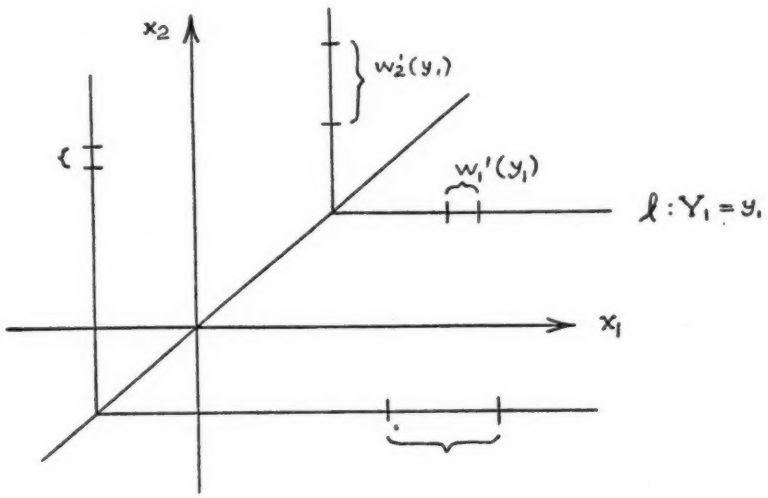


FIG. 1

In general equation (4.9) states that on each hyperplane $Y_1 = y_1$ the relative probability of w is independent of y_1. Since $Y_1 = \min(X_1, \cdots, X_n)$ is a sufficient statistic for θ, Neyman's theorem in this case does give all similar regions.

Next let us consider the case where both extremes of the range of the distribution depend on the parameter. We shall assume (cf. [21]) that $X_1, \cdots, X_n$ are distributed with probability density

$$(4.10)\qquad p(x) = \frac{f(x)}{g(\theta)} \quad \text{in} \quad \theta \leq x \leq b(\theta)$$

where b is a strictly decreasing continuous function over an interval $[-\infty, b(-\infty)]$ and where $b[b(-\infty)] = -\infty$. These assumptions insure that there exists a unique number a, $-\infty < a < b(-\infty)$, such that $b(a) = a$.

Denote by W_{ij}, $(i, j = 1, \cdots, n;\ i \neq j)$, the portion of the sample space where the smallest and the largest of the x's are x_i and x_j respectively. Denote by W_{ij1} and W_{ij2} those portions of W_{ij} where x_i is greater than and less than $b^{-1}(x_j)$ respectively. For any region w denote by w_{ijk}' the intersection of w with

W_{ijk}. Consider a transformation carrying the sample-space into W_{1n}, letting $y_1 = \min(x_1, \cdots, x_n)$, $y_n = \max(x_1, \cdots, x_n)$ and in W_{ij} letting $y_2, \cdots, y_{n-1}$ denote the remaining x's in the order of their subscripts. Next make a transformation carrying W_{1n} into W_{1n1}, letting $z_1 = \max[y_1, b^{-1}(y_n)]$, $z_n = \min[y_1, b^{-1}(y_n)]$ and $z_k = y_k$ for $k = 2, \cdots, n-1$. Denote by w_{ijk} the image of w'_{ijk} in W_{1n1}.

Then Z_n is a sufficient statistic for θ (cf. [21]) and there exist functions f_1, g_1 such that the density of Z_n is given by

$$p(z_n) = \frac{f_1(z_n)}{g_1(\theta)} \quad \text{in} \quad \theta \leq z_n \leq a \tag{4.11}$$

while the distribution of the remaining Z's given Z_n is independent of θ.

The condition that w be a similar region may now be written, analogously to (4.5), in the form

$$\int_a^\theta \frac{f_1(z_n)}{g_1(\theta)} \sum_{i,j,k} \int_{w_{ijk}(z_n)} p(z_1, \cdots, z_{n-1} \mid z_n)\, dz_1 \cdots dz_{n-1}\, dz_n \equiv \epsilon \int_\theta^a \frac{f_1(z_n)}{g_1(\theta)}\, dz_n \tag{4.12}$$

and hence by the argument which led to (4.6), as

$$\sum_{i,j,k} \int_{w_{ijk}(z_n)} p(z_1, \cdots, z_{n-1} \mid z_n)\, dz_1 \cdots dz_{n-1} = \epsilon \qquad \text{for all } z_n. \tag{4.13}$$

Thus in this case also Neyman's theorem gives the most general similar region.

For the case that neither extreme of the range of the distribution depends on the parameter θ, it has been shown by various authors [22, 21, 23] under slightly varying assumptions concerning the regularity of the distribution function, that the existence of a sufficient statistic implies

$$p(x \mid \theta) = \exp[P(\theta) + T(x)Q(\theta) + R(x)]. \tag{4.14}$$

This (cf. [10]) is a special case of that for which Neyman and Pearson determined the totality of similar regions, however under the restriction that the moments of $\Phi = \dfrac{\partial}{\partial\theta} \sum_{i=1}^{n} \log p(X_i)$ uniquely determine the distribution of Φ. We shall briefly indicate how this assumption may be avoided.

Let $X_1, \cdots, X_n$ be a sample from (4.14), or, more generally, (this is the case considered by Neyman and Pearson), let $X_1, \cdots, X_n$ be distributed with probability density

$$p(x_1, \cdots, x_n) = \exp[P(\theta) + u(x_1, \cdots, x_n)Q(\theta) + v(x_1, \cdots, x_n)] \tag{4.15}$$

in a sample space W_+ which is independent of θ. We shall assume that the set of values which Q takes on contains at least some interval. Introducing $\delta =$

$-Q(\theta)$ as a new parameter, we shall obtain all regions similar to δ (where the set of values of δ contains an interval) for the distribution[8]

$$p(x_1, \cdots, x_n) = \exp\,[p_1(\delta) - \delta \cdot u(x_1, \cdots, x_n) + v(x_1, \cdots, x_n)] \tag{4.16}$$

under the assumption that $\sum_{i=1}^{n} \left(\frac{\partial u}{\partial x_i}\right)^2 \neq 0$ except possibly on a set of measure 0.

Let us for a moment assume that there exist functions $f_i(x_i, \cdots, x_n)$, $(i = 2, \cdots, n)$, with continuous partial derivatives almost everywhere and such that the transformation

$$y_1 = u(x_1, \cdots, x_n); \quad y_i = f_i(x_1, \cdots, x_n), \qquad (i = 2, \cdots, n), \tag{4.17}$$

is one to one on W_+ except possibly on a subset of measure 0. Applying this transformation we may write the condition of similarity in the form

$$\int_{-\infty}^{\infty} e^{P_1(\delta)-\delta y_1} \int_{w(y_1)} f(y_1, \cdots, y_n)\, dy_2 \cdots dy_n \cdot dy_1 = \epsilon \int_{-\infty}^{\infty} e^{P_1(\delta)-\delta y_1} \int_{W(y_1)} f(y_1, \cdots, y_n)\, dy_2 \cdots dy_n \cdot dy_1 \tag{4.18}$$

where $W(y_1)$ denotes the region of variation of $y_2, \cdots, y_n$ given y_1, and where $w(y_1)$ denotes the region of variation of $y_2, \cdots, y_n$ given y_1 and w. Furthermore $f(y_1, \cdots, y_n)$ is independent of δ. From the theory of bilateral Laplace transforms it is known that (4.18) implies that

$$\int_{w(y_1)} f(y_1, \cdots, y_n)\, dy_2 \cdots dy_n = \epsilon \int_{W(y_1)} f(y_1, \cdots, y_n)\, dy_2 \cdots dy_n \tag{4.19}$$

which is the desired result.

More generally it may be shown that our assumption concerning $u(x_1, \cdots, x_n)$ insures the existence of functions f_i, $(i = 2, \cdots, n)$, such that under the transformation (4.17) no point $(y_1, \cdots, y_n)$ has more than a denumerable infinity of counter images in x-space. Our proof can be modified to cover this case. The argument is similar to that used to obtain equations (4.9) and (4.13) which were also arrived at through many to one transformations.

5. Testing exponential and rectangular distributions. In their fundamental 1928 paper [24] on likelihood ratio tests, Neyman and Pearson discussed various hypotheses relating to normal, exponential and rectangular distributions. Later they and other authors developed a theory of similar and bisimilar regions which made it possible to obtain optimum tests of many composite hypotheses with

[8] An assumption that we can solve for θ as a function of δ is not needed since we can determine $P_1(\delta)$ by integrating the density (4.16) over W_+.

one constraint concerning normal populations. This theory however is not applicable to most hypotheses concerning exponential or rectangular distributions. We shall in this section obtain optimum tests of some hypotheses relating to these latter distributions, using the method of the previous section.

Let us first consider a sample $X_1, \cdots, X_n$ from an exponential population, the probability density of the sample being.

$$(5.1)\quad p(x_1, \cdots, x_n) = \frac{1}{a^n} \exp\left[-\frac{1}{a}\sum_{i=1}^{n}(x_i - b)\right] \quad \text{if } x_i > b, \quad (i = 1, \cdots, n)$$

and let us consider the two hypotheses $H_1: a =_0$, $H_2: b = b_0$ where, without loss of generality, we shall take $a_0 = 1$, $b_0 = 0$. The likelihood ratio tests of both these hypotheses were shown to be completely unbiased by Paulson [25]. We shall prove

THEOREM 5. *The likelihood ratio tests of H_1 and H_2 are type B_1 and uniformly most powerful, respectively. The one-sided tests based on the likelihood ratio criterion for H_1 are the uniformly most powerful one-sided similar regions for testing this hypothesis.*

PROOF. In order to simplify the argument we shall give a detailed proof only for the restricted class of tests which are symmetric in the variables $X_1, \cdots, X_n$.

For testing H_1 let us make the following transformation introduced by Sukhatme [26]:

$$(5.2)\quad \begin{aligned} Z_1 &= nY_1 \\ Z_i &= (n - i + 1)(Y_i - Y_{i-1}), \qquad (i = 2, \cdots, n), \end{aligned}$$

where Y_i is the ith of the X's in order of magnitude. Then

$$(5.3)\quad p(z_1, \cdots, z_n) = \frac{1}{a^n} \exp\left[-\frac{1}{a}(z_1 - nb) - \frac{1}{a}\sum_{i=2}^{n} z_i\right] \quad \text{if} \quad z_1 \geq nb; z_i \geq 0 \qquad (i = 2, \cdots, n).$$

We want to determine all regions w which under H are similar to the sample space with respect to b, i.e. all regions w satisfying

$$(5.4)\quad \begin{aligned} &\int_w e^{-(z_1 - nb)} \exp\left[-\sum_{i=2}^{n} z_i\right] dz_2 \cdots dz_n\, dz_1 \\ &\qquad = \int_{nb}^{\infty} e^{-(z_1 - nb)} \left\{\int_{w(z_1)} \exp\left[-\sum_{i=2}^{n} z_i\right] dz_2 \cdots dz_n\right\} dz_1 \\ &\qquad \overset{(b)}{\equiv} \epsilon \overset{(b)}{\equiv} \epsilon \int_{nb}^{\infty} e^{-(z_1 - nb)}\, dz_1 \end{aligned}$$

where $w(z_1)$ denotes the intersection of w with the hyperplane $Z_1 = z_1$. Now (5.4) is equivalent to

$$e^{nb}\int_{nb}^{\infty} e^{-z_1} f(z_1)\, dz_1 \overset{(b)}{\equiv} 0 \tag{5.5}$$

where

$$f(z_1) = \int_{w(z_1)} \exp\left[-\sum_{i=2}^{n} z_i\right] dz_2 \cdots dz_n - \epsilon \tag{5.6}$$

and this in turn is equivalent to

$$f(z_1) = 0 \text{ for all } z_1 . \tag{5.7}$$

Of all the regions w satisfying (5.7) we want to determine the one which against a specific alternative, say a_1, has maximum power, i.e. for which

$$\int_{nb}^{\infty} e^{-(1/a_1)(z_1 - nb)} \int_{w(z_1)} \exp\left[-\frac{1}{a_1}\sum_{i=2}^{n} z_i\right] dz_2 \cdots dz_n\, dz_1 \tag{5.8}$$

is as large as possible. We thus see that w will have the desired properties if $w(z_1)$ is determined according to the two conditions

$$\int_{w(z_1)} \exp\left[-\sum_{i=2}^{n} z_i\right] dz_2 \cdots dz_n = \epsilon \tag{5.9}$$

and

$$\int_{w(z_1)} \exp\left[-\frac{1}{a_1}\sum_{i=2}^{n} z_i\right] dz_2 \cdots dz_n = \max. \tag{5.10}$$

Hence by the Neyman-Pearson fundamental lemma $w(z_1)$ is the set of points satisfying

$$\exp\left[\left(-\frac{1}{a_1}\sum_{i=2}^{n} z_i + \sum_{i=2}^{n} z_i\right)\right] \geq C(a_1, z_1) \tag{5.11}$$

and therefore according as a_1 is greater or less than 1, $w(z_1)$ is determined by

$$\begin{aligned} \sum_{i=2}^{n} z_i &= \sum_{i=1}^{n} [x_i - \min(x_1, \cdots, x_n)] \geq k(a_1, z_1), \text{ or} \\ \sum_{i=2}^{n} z_i &= \sum_{i=1}^{n} [x_i - \min(x_1, \cdots, x_n)] \leq k'(a_1, z_1). \end{aligned} \tag{5.12}$$

But $\sum_{i=2}^{n} Z_i$ is independently distributed of Z_1 and under H the distribution of $\sum_{i=2}^{n} Z_i$ does not depend on a_1, in fact it is a chi-square distribution with $2n - 2$ degrees of freedom. Thus k and k', as determined by (5.9) are independent of a_1 and the two tests (5.12) are uniformly most powerful one-sided.

Next we consider the more restricted class of unbiased similar regions. For w to be unbiased we must have

$$\frac{d}{da}\left\{\frac{1}{a^n}\int_w \exp\left[-\frac{z_1-nb}{a}\right]\exp\left[-\frac{1}{a}\sum_{i=2}^n z_i\right]dz_1\cdots dz_n\right\}\Big|_{a=1}$$

$$(5.13)\quad = \int_{nb}^\infty (z_1-nb-n)\exp\left[-(z_1-nb)\right]\int_{w(z_1)}\exp\left[-\sum_{i=2}^n z_i\right]dz_2\cdots dz_n\,dz_1$$

$$+\int_{nb}^\infty \exp\left[-(z_1-nb)\right]\int_{w(z_1)}\left(\sum_{i=2}^n z_i\right)\exp\left[-\sum_{i=2}^n z_i\right]dz_2\cdots dz_n\,dz_1 = 0.$$

The first of the integrals in the middle member equals

$$(5.14)\quad \begin{aligned}&\int_0^\infty (z-n)\,e^{-z}\int_{w(z+nb)}\exp\left[-\sum_{i=2}^n z_i\right]dz_2\cdots dz_n\,dz\\ &= \epsilon\int_0^\infty (z-n)\,e^{-z}\,dz = -(n-1)\epsilon.\end{aligned}$$

Therefore

$$(5.15)\quad \begin{aligned}&\int_{nb}^\infty e^{-(z_1-nb)}\int_{w(z_1)}\left(\sum_{i=2}^n z_i\right)\exp\left[-\sum_{i=2}^n z_i\right]dz_2\cdots dz_n\,dz_1\\ &= (n-1)\epsilon = (n-1)\epsilon\int_{nb}^\infty e^{-(z_1-nb)}\,dz_1\end{aligned}$$

or

$$(5.16)\quad \int_{nb}^\infty e^{-z_1}\,g(z_1)\,dz_1 \overset{(b)}{\equiv} 0$$

where

$$(5.17)\quad g(z_1) = \int_{w(z_1)}\left(\sum_{i=2}^n z_i\right)\exp\left[-\sum_{i=2}^n z_i\right]dz_2\cdots dz_n - (n-1)\epsilon.$$

Thus finally the condition of unbiasedness reduces to

$$(5.18)\quad \int_{w(z_1)}\left(\sum_{i=2}^n z_i\right)\exp\left[-\sum_{i=2}^n z_i\right]dz_2\cdots dz_n = (n-1)\epsilon$$

and we seek the region $w(z_1)$ which satisfies (5.9), (5.10) and (5.18).

By the fundamental lemma $w(z_1)$ is given by

$$(5.19)\quad \exp\left[-\frac{1}{a_1}\sum_{i=2}^n z_i\right] \geq \left[C_1(a_1, z_1)\sum_{i=2}^n z_i + C_2(a_1, z_1)\right]\cdot\exp\left[-\sum_{i=2}^n z_i\right]$$

which is equivalent to

$$(5.20)\quad \sum_{i=2}^n z_i \leq k_1(a_1, z_1),\ \geq k_2(a_1, z_1)$$

where k_1 and k_2 are determined by (5.9) and (5.18), and are therefore independent of z_1 and a. Thus the region (5.20) which of all unbiased similar regions

maximizes the power against the alternative $a = a_1$ is independent of a_1 and hence is a region of type B_1. This completes the proof since it is easily verified that (5.10) is equivalent to the likelihood ratio test.

The proof for regions which are not necessarily symmetric in the variables follows similarly if instead of the transformation (5.2) one uses a transformation $U_i = f_i(X_1, \cdots, X_n)$ which is one to one and such that $U_1 = Z_1$, $U_2 = \sum_{i=2}^{n} Z_i$. The distribution of $U_3, \cdots, U_n$ is then independent of a and b since U_1, U_2 are a pair of sufficient statistics for these parameters, and the proof carries over step by step.

Next we consider the hypothesis $H_2: b = 0$, and again we restrict ourselves to regions which are symmetric in the variables, although as before the proof can be modified to cover also nonsymmetric regions.

We first make the transformation to $Z_1, \cdots, Z_n$ given by (5.2). In the $n - 1$ dimensional space of $Z_2, \cdots, Z_n$, we then transform to new variables $U, \Psi_1, \cdots, \Psi_{n-2}$ where $U = \sum_{i=2}^{n} Z_i$ and where the Ψ's are the generalized polar angles. Obviously the distribution of the Ψ's does not depend on a, since they are homogeneous of degree 0 in the Z's. Furthermore the Ψ's are independently distributed of U since the probability density of the Z's is constant over the hyperplanes $U = u$. Thus

$$p(z_1, u, \psi_1, \cdots, \psi_{n-2}) = \frac{K}{a^n} \exp\left[-\frac{z_1 - nb}{a}\right] u^{n-2} e^{-u/a} p(\psi_1, \cdots, \psi_{n-2}). \tag{5.21}$$

We next introduce new variables

$$V = Z_1 + U \text{ and } T = \frac{Z_1}{Z_1 + U} \tag{5.22}$$

and find

$$p(v, t, \psi_1, \cdots, \psi_{n-2}) = \frac{K}{a^n} \exp\left[-\frac{v - nb}{a}\right] v^{n-1} (1 - t)^{n-2} p(\psi_1, \cdots, \psi_{n-2}) \tag{5.23}$$

$$\text{for } v \geq nb, \frac{nb}{v} \leq t \leq 1.$$

For w under H_2 to be similar with respect to a, we must have

$$\int_0^\infty \frac{K}{a^n} \exp\left[-\frac{v}{a}\right] v^{n-1} \int_{w_0(v)} (1 - t)^{n-2} p(\psi_1, \cdots, \psi_{n-2})\, dt\, d\psi_1 \cdots d\psi_{n-2} \cdot dv = \epsilon \int_0^\infty \frac{K}{a^n} \exp\left[-\frac{v}{a}\right] v^{n-1}\, dv \tag{5.24}$$

where $w(v)$ designates the intersection of w with the hyperplane $V = v$, and where $w_0(v)$ denotes the part of $w(v)$ lying between the hyperplanes $t = 0$ and $t = 1$.

Hence the condition of similarity may be written as

$$\int_0^\infty \exp\left[-\frac{v}{a}\right] v^{n-1} f(v)\, dv = 0 \qquad \text{for all } a > 0 \tag{5.25}$$

where

$$f(v) = \int_{w_0(v)} (1-t)^{n-2} p(\psi_1, \cdots \psi_{n-2})\, dt\, d\psi_1 \cdots d\psi_{n-2} - \epsilon. \tag{5.26}$$

By the uniqueness theorem for Laplace transforms, (5.25) implies $f(v) = 0$ for all $v > 0$, so that the condition of similarity finally reduces to

$$\int_{w_0(v)} (1-t)^{n-2} p(\psi_1, \cdots, \psi_{n-2})\, dt\, d\psi_1 \cdots d\psi_{n-2} = \epsilon. \tag{5.27}$$

Of all similar regions, let us find the one which has maximum power. Obviously we want to include in $w(v)$ all points for which $t < 0$. In addition we want to choose $w_0(v)$ such that

$$\int_{w_b(v)} (1-t)^{n-2} p(\psi_1, \cdots, \psi_{n-2})\, dt\, d\psi_1 \cdots d\psi_{n-2} = \max \tag{5.28}$$

where $w_b(v)$ is that part of $w(v)$ in which $\max\left(0, \frac{nb}{v}\right) \leq t$.

If, for some alternative b, $w_0(v)$ is contained in $\frac{nb}{v} < t < 1$, then $w_b(v)$ and $w_0(v)$ coincide and hence (5.28) attains its maximum value ϵ whatever the position of $w_0(v)$ in $\frac{nb}{v} \leq t \leq 1$. If on the other hand $\frac{nb}{v}$ is so close to 1 that $\frac{nb}{v} \leq t \leq 1$ is too small to contain $w_0(v)$, then (5.28) attains its maximum for any $w_0(v)$ containing $\frac{nb}{v} \leq t \leq 1$. There exists therefore a unique $w_0(v)$ which maximizes (5.28) for all values of b and v, namely the region defined by

$$C(v) \leq t \leq 1 \tag{5.29}$$

where C is determined by (5.27).

Since under H_2, the statistics V and T are independent, C does not depend on v. The test

$$t \leq 0, \qquad \geq C \tag{5.30}$$

which we have just shown to be uniformly most powerful, is also the likelihood ratio test which completes the proof of the theorem.

We shall finally consider an example of an optimum test in connection with a

rectangular distribution. Let $X_1, \cdots, X_n$ be independently and uniformly distributed over $(a, a + \theta)$, where θ is positive. *For testing the hypothesis* $H: a = a_0$, *the test*

$$\frac{Y_1 - a_0}{Y_n - Y_1} \leq 0, \qquad \geq C \tag{5.31}$$

where Y_1 *and* Y_n *are the smallest and the largest of the* X's *respectively, is the uniformly most powerful of all similar regions.*

The proof of this goes through very much like that for H_2 in Theorem 5. Without loss of generality we take $a_0 = 0$. Also again, to simplify the proof, we restrict ourselves to regions which are symmetric in the variables. We need the following lemma.

LEMMA. *Let* $X_1, \cdots, X_n$ *be independently and uniformly distributed over* $(a, a + \theta)$. *Let* Y_i *denote the* i*th* X *in order of magnitude, and let*

$$T_n = Y_n, \; T_k = \frac{Y_k}{Y_{k+1}}, \qquad (= 1, \cdots, n - 1). \tag{5.32}$$

Then for $a > 0$

$$p(t_1, \cdots, t_n) = \frac{n!}{\theta^n} t_n^{n-1} t_{n-1}^{n-2} \cdots t_2 \tag{5.33}$$

when

$$a \leq t_n \leq a + \theta, \; \frac{a}{t_n \cdot t_{n-1} \cdots t_{k+1}} \leq t_k \leq 1, \qquad (k = 1, \cdots, n - 1).$$

This is easily seen by applying the usual method of Jacobians. The inequalities describing the sample space of the T's are equivalent to the following more convenient ones:

$$a \leq t_n \leq a + \theta, \; \frac{a}{t_n} \leq t_1 t_2 \cdots t_{n-1} \leq 1; t_k \leq 1, \qquad (k = 1, \cdots, n - 1). \tag{5.34}$$

Let us denote by $w(t_n)$ the intersection of a region w with the hyperplane $T_n = t_n$, and by $w_0(t_n)$ that part of $w(t_n)$ contained in the cylinder $0 \leq t_k \leq 1$, $(k = 1, \cdots, n - 1)$; then we find as a necessary and sufficient condition for w to be similar with respect to θ (assuming H)

$$(n - 1)! \int_{w_0(t_n)} t_{n-1}^{n-2} t_{n-2}^{n-3} \cdots t_2 \, dt_{n-1} \cdots dt_1 = \epsilon. \tag{5.35}$$

Of all regions satisfying (5.35) we want to find the most powerful one. Let us first consider alternatives $a > 0$. If $w_a(t_n)$ denotes the common part of $w_0(t_n)$ and the region

$$\frac{a}{t_n} \leq t_{n-1} t_{n-2} \cdots t_1 \leq 1, \tag{5.36}$$

we must choose $w_a(t_n)$ such that

$$\int_{w_0(t_n)} t_{n-1}^{n-2} \cdot t_{n-2}^{n-3} \cdots t_2 \, dt_{n-1} \cdots dt_1 = \max. \tag{5.37}$$

From this it follows easily that against alternatives $a > 0$ the uniformly best choice for $w_0(t_n)$ is

$$t_1 t_2 \cdots t_{n-1} = \frac{y_1}{y_n} \geq C'(t_n), \tag{5.38}$$

and since under H, $\frac{Y_1}{Y_n}$ is independently distributed of T_n, $C'(t_n)$ does not depend on t_n.

Consider next alternatives $a < 0$. We include in the region of rejection all points for which $Y_1 \leq 0$. To determine $w_0(t_n)$ we notice that, given $Y_1 > 0$, the X's are uniformly distributed between 0 and $a + \theta$. (Provided $a + \theta > 0$; the case $a + \theta \leq 0$ is trivial). Hence the probability distribution of the T's given $Y_1 > 0$ is

$$p(t_1, \cdots, t_n \mid Y_1 > 0) = \frac{n!}{(a+\theta)^n} t_n^{n-1} \cdots t_2 \tag{5.39}$$

when

$$0 \leq t_n \leq a + \theta, \qquad 0 \leq t_k \leq 1 \qquad \text{for } k = 1, \cdots, n-1.$$

Thus

$$\frac{p(t_1, \cdots, t_{n-1} \mid t_n, a < 0, Y_1 > 0)}{p(t_1, \cdots, t_{n-1} \mid t_n, a = 0)} \tag{5.40}$$

is independent of $t_1, \cdots, t_{n-1}$ and hence the power of w against alternatives $a < 0$ is independent of the choice of $w_0(t_n)$. Therefore the region

$$y_1 \leq 0, \ \frac{y_1}{y_n} \geq C' \tag{5.41}$$

is uniformly most powerful against all alternatives. But (5.41) is equivalent to

$$\frac{y_1}{y_n - y_1} \leq 0, \geq C. \tag{5.42}$$

It is interesting to compare this result with that for the corresponding simple hypothesis. Let H' be the hypothesis: $a = 0$ when the X's are assumed independently and uniformly distributed over $(a, a + 1)$. There exists no uniformly most powerful test of H'; instead the two uniformly most powerful one-sided tests exist. By analogy with the normal case one might then expect for H' that of all tests with symmetric power-functions, there be a uniformly most powerful one. This however is not so: there exist infinitely many admissible tests with symmetric powerfunction.

In this and the previous section we restricted ourselves to problems involving only one nuisance parameter. However, the method applies also to problems involving several nuisance parameters.

In the usual way (cf. [20, 9]) the results of this section may be translated to give optimum sets of confidence intervals for estimating the parameters in question. In this connection it is an open question whether the confidence regions based on the type B_1 tests discussed in section 2 will always be intervals; one would expect this to be the case.

The author wishes to acknowledge his indebtedness to Professor P. L. Hsu for many helpful suggestions.

Added in proof: In a joint paper by Professor Henry Scheffé and the present author which has been submitted to the Proceedings of the National Academy of Sciences, a result is given concerning the existence of certain 1:1 transformations. This result bears on Section 4 of the present paper where a question arises concerning the existence of a 1:1 transformation. The existence of such a transformation is now assured and, as a consequence, the last paragraph of Section 4 has become superfluous.

REFERENCES

[1] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses", *Roy. Soc. London Phil. Trans.*, Ser. A, Vol. 231 (1933), p. 289.

[2] J. Neyman, "Sur la vérification des hypothèses statistiques composées", *Bull. Soc. Math. France*, Vol. 63 (1935), p. 1.

[3] J. Neyman, "On a statistical problem arising in routine analysis and in sampling inspection of mass production", *Annals of Math. Stat.*, Vol. 12 (1941), p. 46.

[4] H. Scheffé, "On the theory of testing composite hypotheses with one constraint", *Annals of Math. Stat.*, Vol. 13 (1942), p. 280.

[5] P. L. Hsu, "Analysis of variance from the power function standpoint", *Biometrika*, Vol. 32 (1941), p. 62.

[6] J. B. Simaika, "On an optimum property of two important statistical tests", *Biometrika*, Vol. 32 (1941), p. 70.

[7] A. Wald, "On the power function of the analysis of variance test", *Annals of Math. Stat.*, Vol. 13 (1942), p. 434.

[8] P. L. Hsu, "On the power function of the E^2-test and the T^2-test", *Annals of Math. Stat.*, Vol. 16 (1945), p. 278.

[9] H. Scheffé, "On the ratio of the variances of two normal populations", *Annals of Math. Stat.*, Vol. 13 (1942), p. 371.

[10] J. Neyman, "L'estimation statistique traitée comme un problème classique de probabilité", *Actualités Sci. et Ind. No. 739*, Hermann et Cie, Paris, 1938.

[11] A. Wald, "Notes on the theory of statistical estimation and testing hypotheses", Columbia University.

[12] W. J. Dixon, "Further contributions to the problem of serial correlation", *Annals of Math. Stat.*, Vol. 15 (1944), p. 119.

[13] J. von Neumann, "Distribution of the ratio of the mean square successive difference to the variance", *Annals of Math. Stat.*, Vol. 12 (1941), p. 367.

[14] M. Kac, "A remark on independence of linear and quadratic forms involving independent Gaussian variables", *Annals of Math. Stat.*, Vol. 16 (1945), p. 400.

[15] W. G. Madow, "Note on the distribution of the serial correlation coefficient", *Annals of Math. Stat.*, Vol. 16 (1945), p. 308.

[16] R. L. ANDERSON, "Distribution of the serial correlation coefficient", *Annals of Math. Stat.*, Vol. 13 (1942), p. 1.

[17] T. KOOPMANS, "Serial correlation and quadratic forms in normal variables", *Annals of Math. Stat.*, Vol. 13 (1942), p. 14.

[18] R. B. LEIPNIK, "Distribution of the serial correlation coefficient in a circularly correlated universe", *Annals of Math. Stat.*, Vol. 18 (1947), p. 80.

[19] P. L. HSU, "On the asymptotic distributions of certain statistics used in testing the independence between successive observations from a normal population", *Annals of Math. Stat.*, Vol. 17 (1946), p. 350.

[20] J. NEYMAN, "Outline of a theory of statistical estimation based on the classical theory of probability", *Roy. Soc. London Phil. Trans.*, Ser. A, Vol. 236 (1937), p. 333.

[21] E. J. G. PITMAN, "Sufficient statistics and intrinsic accuracy", *Proc. Camb. Phil. Soc.*, Vol. 32 (1936), p. 567.

[22] B. O. KOOPMAN, "On distributions admitting a sufficient statistic", *Trans. Amer. Math. Soc.*, Vol. 39 (1936), p. 399.

[23] D. DUGUÉ, "Application des propriétés de la limite au sens du calcul des probabilités à l'étude des diverses questions d'estimation", *J. École Poly.*, Vol. 3 (1937), p. 305.

[24] J. NEYMAN AND E. S. PEARSON, "On the use and interpretation of certain test criteria for purposes of statistical inference", *Biometrika*, Vol. 20A (1928), p. 175.

[25] E. PAULSON, "On certain likelihood ratio tests associated with the exponential distribution", *Annals of Math Stat.*, Vol. 12 (1941), p. 301.

[26] P. V. SUKHATME, "Tests of significance for samples of the χ^2 population with two degrees of freedom", *Annals of Eugenics*, Vol. 8 (1937), p. 52.

A CORNER TEST FOR ASSOCIATION

By Paul S. Olmstead and John W. Tukey

Bell Telephone Laboratories and Princeton University

1. Summary. This paper proposes a new test (the "quadrant sum") for the association of two continuous variables. Its notable properties are:

(1) Special weight is given to extreme values of the variables.

(2) Computation is very easy.

(3) The test is non-parametric.

Significance levels (for the quadrant sum) are given to the accuracy needed for practical use. To this accuracy they are independent of sample size (see Fig. 1). The generating function of the quadrant sum is given for the null hypothesis (no association = independence). A limiting distribution is deduced and compared with the cases $2n = 4, 6, 8, 10$, and 14. Extension to higher dimensions and application to serial correlation are discussed.

2. Description of test (even number in sample). We shall describe the test as though a scatter diagram had already been drawn. The possibilities of direct computation from tabular data are indicated by the examples in sections 8 and 9.

In the scatter diagram, draw the two lines, $x = x_m$, $y = y_m$, where x_m is the median of the x-values without regard to the values of y, and y_m is the median of the y-values without regard to the values of x. Think of the four quadrants or corners thus formed as being labelled $+, -, +, -$, in order, so that the upper right and lower left quadrants are positive. Beginning at the right hand side of the diagram, count in (in order of abscissae) along the observations until forced to cross the horizontal median. Write down the number of observations met before this crossing, attaching the sign $+$ if they lay in the $+$ quadrant, and the sign $-$ if they lay in the $-$ quadrant. Repeat this process moving up from below, moving to the right from the left, and moving down from above. The quadrant sum is the algebraic sum of the four terms thus written down. This process is illustrated in Fig. 2, where the black dots represent contributions to the sum, and the dotted lines, crossings.

When there are an even number of pairs (x, y) and no ties, the medians will pass between the points. In this, the simplest case, the distribution of the quadrant sum is known for the hypothesis of no association (that is, of independence), and significance levels are given in Table 1 for the magnitude (absolute value) of the sum. It will be noticed that the sample size does not enter in any important way.

The cases of an odd number of observations and of ties are discussed in the next two sections. Simple devices make the test usable in most cases. A very great tendency toward ties, however, will make it inapplicable. This will be

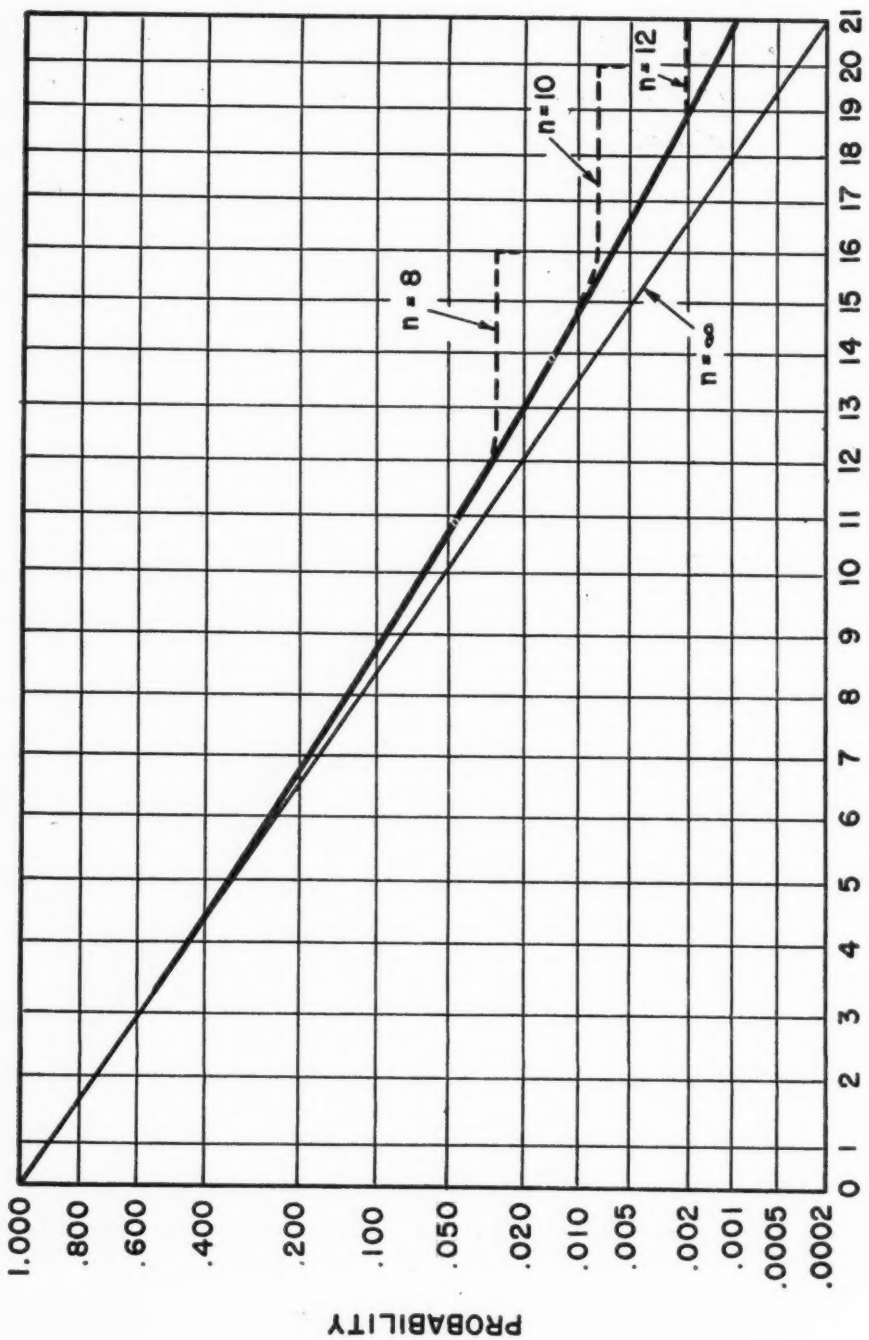


FIG. 1. Probability that an arrangement of a sample of size, N, will have a quadrant sum equal to or greater than indicated absolute value.

unimportant in most applications because of the fact that attention is being directed to the periphery.

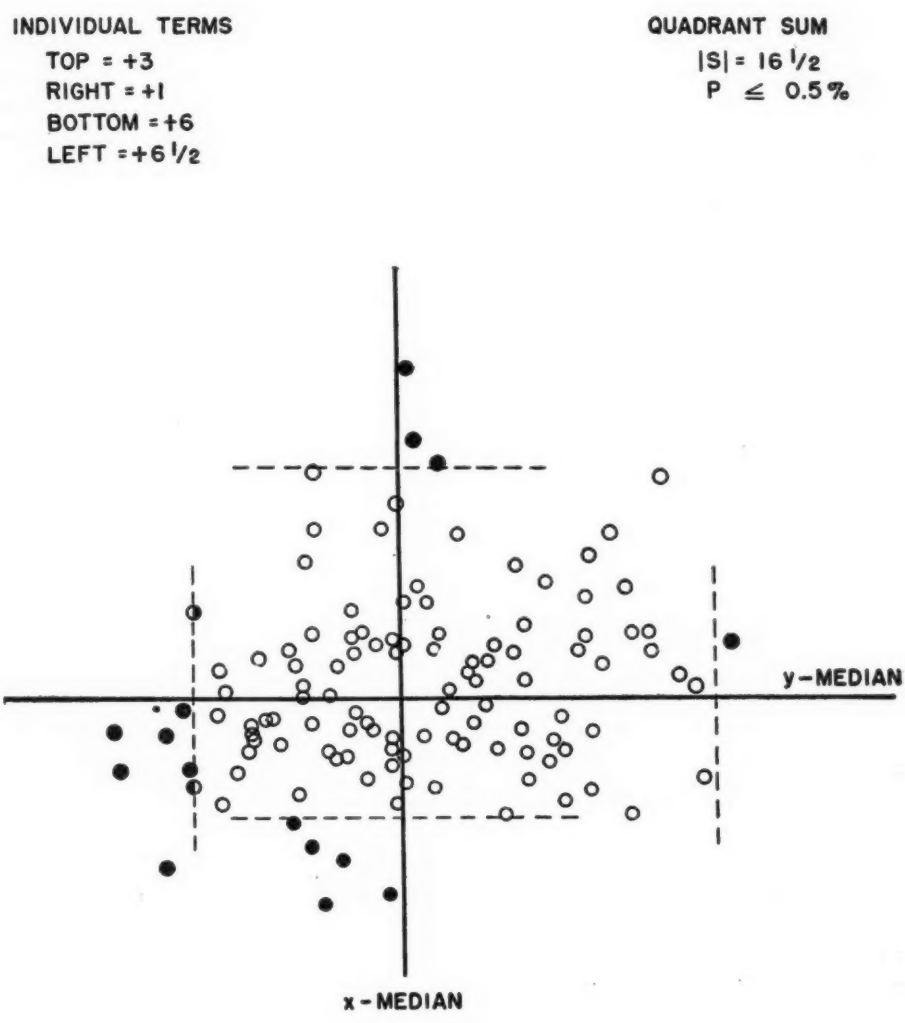


FIG. 2. Scatter diagram of 116 pairs of observations

The set of data which prompted the development of the test is shown in Fig. 2. The accompanying report described it as follows: "The various points appear to be scattered almost completely at random and give little indication of correlation." The quadrant sum is $16\frac{1}{2}$ which is significant at the 0.5% point. Intuitively, the significant association of the peripheral points is clear.

3. Description of test (odd number in sample). If the sample size is odd, then we may usually follow the process outlined above. We will have difficulty only when the counting process meets a point, one of whose coordinates is a median. In this case we employ a simple device, namely:

Given a sample of $2n + 1$ pairs, let x^* and y^* be the medians of the x-values and of the y-values, respectively. Let the pairs in which they occur be (x^*, y_k) and (x_m, y^*), respectively. Replace these two pairs by the sing'e pair (x_m, y_k). There are now $2n$ pairs and the regular method can be applied.

The quadrant sum so obtained from an unassociated population has the same distribution as that formed directly from $2n$ pairs.

4. Description of test (treatment of ties). The behavior of the test is known when (1) there is no association, (2) the probability of a tie in x-values or y-values is zero. The following approximation, which has an unknown effect on the distribution, is suggested when ties are present:

TABLE 1

Working significance levels for magnitudes of quadrant sums

Significance level (Conservative)	Magnitude of quadrant sum*
10%	9
5%	11
2%	13
1%	14–15
0.5%	15–17
0.2%	17–19
0.1%	18–21

* The smaller magnitude applies for large sample size, the larger magnitude for small sample size. Magnitudes equal to or greater than twice the sample size less six should not be used.

When a tied group is reached, count the number in the tied group favorable to continuing and the number unfavorable. Treat the tied group as if the number of its points preceding the crossing of the median were

$$\frac{\text{number favorable}}{1 + \text{number unfavorable}}.$$

It seems likely that this approximation is conservative.

5. Discussion. When a moderate number, say 25 to 200, of paired observations on two quantities are plotted as a scatter diagram, visual examination frequently detects what seems to be definite evidence of association between the variables. Often in such cases, the usual methods for measuring associa-

tion do not find statistical significance of association. Visual judgment, particularly by engineers or scientists who may wish to take action on the basis of their findings, gives greater weight to observations near the periphery of the scatter diagram. This is not always desirable—but often it is very desirable. A quantitative test of association with such concentration on the periphery has been lacking. The quadrant sum test was developed to fill the gap. Its features of speed and non-parametricity are useful but secondary from this point of view.

When uniform attention to the whole scatter diagram is desired, the quadrant sum test is of unknown usefulness. We know little enough of the operating characteristics of the more conventional tests, such as:

1. The product moment correlation coefficient
2. The four-fold table formed by the medians
3. The biserial correlation coefficient
4. The rank correlation coefficient

and less about the operating characteristics of the present test. In this case, the quadrant sum test can only be recommended definitely for exploratory investigations of large amounts of data.

There are many situations, however, where we do not know where to concentrate our attention, and where speed and non-parametricity are cardinal virtues in a test. One example is the use of serial correlation in studying industrial processes. We may guess that here we are interested in the periphery, but neither theory nor experience can, so far, prove this. In such situations the quadrant sum is by far the fastest to use of any of the tests known to the authors, and we believe one of the most useful.

6. Elementary derivations. We can easily find the distribution of

1. An individual term of the quadrant sum
   a. For fixed sample size
   b. In the limit
2. The quadrant sum itself
   a. For fixed sample size
   b. In the limit, assuming asymptotic independence of the four terms.

This we shall do now, leaving the proof that 2a actually converges to 2b to a later section.

Consider a sample of $2n$ pairs $(x_1, y_1), \cdots, (x_{2n}, y_{2n})$ from a population in which x and y are independent. It is both clear and easily verifiable that

1. The set of $2n$ x-values, $x_1, \cdots, x_{2n}$
2. The set of $2n$ y-values, $y_1, \cdots, y_{2n}$
3. The permutation of the order of the y-values when the pairs are ordered by the x-values

which together determine the sample, are independently distributed, and that any permutation is as likely as every other. (We have assumed no ties, which is a consequence, with probability one, of the continuous cumulative distribu-

tions of x and y). Since the quadrant sum depends only on the permutation, its distribution in the absence of association does not depend on the distributions of x and y.

We must solve, then, certain purely combinatorial problems—under the hypothesis that the $2n!$ permutations of the y-values are all equally likely. It may simplify matters to assume that the values of x in the sample are $1, 2, \cdots, 2n$ and that those of y are the same. How, then, do we calculate the distribution of a single term of the quadrant sum. Let us begin with small x-values, and the pair $(1, y_1)$. If $y_1 = 1, 2, \cdots, n$, we count "one" positive, and if $y_1 = n + 1, n + 2, \cdots, 2n$, we count "one" negative. We pass on to $(2, y_2)$ and so on. How many permutations yield a count of exactly k positive values? Those in which $y_1, y_2, \cdots, y_k$ are equal to or less than n, y_{k+1} equal to or greater than $n + 1$, and the other $(2n - k - 1)y$'s are arbitrary. There are:

$$n(n - 1) \cdots (n - k + 1) \cdot (n)(2n - k - 1)!$$

such permutations, the fraction of all $(2n)!$ permutations being:

$$\frac{n(n - 1) \cdots (n - k + 1)n}{(2n)(2n - 1) \cdots (2n - k + 1)(2n - k)} \tag{1}$$

which is, then, the probability that this contribution will equal $+k$, or by symmetry, the probability that it will equal $-k$, $k \neq 0$.

For large n, this becomes merely:

$$p_k = 2^{-(|k|+1)}, \qquad k \neq 0. \tag{2}$$

In order to obtain the distribution of the quadrant sum itself, we must concern ourselves with the lack of independence of the four terms. This is indicated most clearly in the case of $2n = 2$, where the $2! = 2$ permutations yield $+1 +1 +1 +1 = 4$ and $-1 -1 -1 -1 = -4$. Here, there is complete lack of independence. We shall see later that there is effectively independence in the limit, so that it is worth while to calculate the sum of four independent terms with the limiting distribution (2) and find that it satisfies:

$$Pr(|\text{ independent sum of 4 terms }| \geq k) = \frac{9k^3 + 9k^2 + 168k + 208}{216 \cdot 2^k}, k > 0. \tag{3}$$

The details will be omitted.

A simple device, reminiscent of Wald's [3, 1943] establishment of the two-dimensional tolerance limits enables us to avoid difficulties with lack of independence and compute the exact distribution of the quadrant sum for any n. We decompose the permutation of the y-values into the following parts, which together specify the permutation:

(a) The number, j, of pairs in the upper right quadrant.

(b) The set of j values of x between $n + 1$ and $2n$ corresponding to pairs in the upper right quadrant.

(c) The set of j values of y between $n + 1$ and $2n$ corresponding to points in the upper right quadrant.
(d) The set of j values of x between 1 and n corresponding to pairs in the lower left quadrant. (Note that the use of medians ensures that the lower left and upper right quadrants contain the same number of points.)
(e) The set of j values of y between 1 and n corresponding to pairs in the lower left quadrant.
(f) The permutation of j objects defined by the pairs in the upper right quadrant.
(g) The permutation of $n - j$ objects defined by the pairs in the upper left quadrant.
(h) and (i) the permutations from the remaining quadrants.

It is easily verified that: (1) given j, items (b) to (i) can be assigned at will, (2) each assignment of (a) to (i) corresponds to one and only one permutation, (3) the quadrant sum depends only on items (b) to (e). In fact, the right hand term depends on item (b), the upper term on item (c), the left hand term on item (d) and the lower term on item (e). While j remains fixed, the terms behave independently.

For fixed j, what is the distribution of a single term? If a set of j x-values gives the term $+k$, it must contain the k largest x-values and not contain the next. There are:

$$\binom{n-k-1}{n-j-1}$$

such sets. The generating function for a single term, is, then:

$$\sum_{k=1}^{j}\binom{n-k-1}{n-j-1}x^k + \sum_{k=1}^{n-j}\binom{n-k-1}{j-1}x^{-k}. \tag{4}$$

Since the terms are independent for fixed j, and there are $(j!)^2((n-j)!)^2$ ways to supply the permutations forming items (f) to (i), the generating function for the quadrant sum, S_n, is:

$$G_n(x) = \sum_{j=0}^{n}\frac{(j!)^2((n-j)!)^2}{(2n)!}\left[\sum_{k=1}^{j}\binom{n-k-1}{n-j-1}x^k + \sum_{k=1}^{n-j}\binom{n-k-1}{j-1}x^{-k}\right]^4. \tag{5}$$

The exact probability of equalling or exceeding each value of S_n has been computed for $2n = 2, 4, 6, 8, 10$, and 14. Table 2 gives these probabilities and Fig. 3 shows the values of

$$\frac{m}{5} + \log_{10} Pr(\,|\text{ quadrant sum }| \geq m)$$

this particular function being chosen for its relative constancy. The maximum value of the quadrant sum is $4n$, and for values of k less than $4n - 6$, there

TABLE 2

Probability of a Sum of Absolute Value Equal to or Greater than k when a Sample of 2n is Drawn from an Unassociated Population

$\lvert k\rvert$ \ $2n$	2	4	6	8	10	12	14	∞*
0	1.0000	1.0000	1.0000	1.0000	1.0000		1.0000	1.000000
1	1.0000	0.7500	0.9333	0.9036	0.9106		0.9115	0.912037
2	1.0000	0.7500	0.7556	0.7544	0.7567		0.7580	0.754630
3	1.0000	0.4167	0.6000	0.6000	0.6008		0.6039	0.599537
4	1.0000	0.4167	0.4667	0.4619	0.4662		0.4690	0.462963
5	0.0000	0.3333	0.3111	0.3508	0.3519		0.3547	0.346933
6	0.0000	0.3333	0.2222	0.2619	0.2589		0.2611	0.252025
7	0.0000	0.3333	0.1556	0.1821	0.1867		0.1876	0.177662
8	0.0000	0.3333	0.1111	0.1258	0.1333		0.1322	0.121817
9	0.0000	0.0000	0.1000	0.0839	0.0928		0.0918	0.081471
10	0.0000	0.0000	0.1000	0.0554	0.0642		0.0632	0.053295
11	0.0000	0.0000	0.1000	0.0375	0.0436		0.0432	0.034189
12	0.0000	0.0000	0.1000	0.0304	0.0290		0.0296	0.021557
13	0.0000	0.0000	0.0000	0.0286	0.0190		0.0202	0.013386
14	0.0000	0.0000	0.0000	0.0286	0.0127		0.0139	0.008200
15	0.0000	0.0000	0.0000	0.0286	0.0095		0.0096	0.004963
16	0.0000	0.0000	0.0000	0.0286	0.0083		0.0066	0.002972
17	0.0000	0.0000	0.0000	0.0000	0.0079		0.0045	0.001762
18	0.0000	0.0000	0.0000	0.0000	0.0079		0.0031	0.001036
19	0.0000	0.0000	0.0000	0.0000	0.0079		0.0021	0.000604
20	0.0000	0.0000	0.0000	0.0000	0.0079		0.0014	0.000350
21	0.0000	0.0000	0.0000	0.0000	0.0000		0.0010	0.000201
22	0.0000	0.0000	0.0000	0.0000	0.0000		0.0008	0.000115
23	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000065
24	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000036
25	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000020
26	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000011
27	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000006
28	0.0000	0.0000	0.0000	0.0000	0.0000		0.0006	0.000003
29	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.000002
30	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.000001
31 or over	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.000000
Variance of k	16	24	$26\frac{2}{5}$	$26\frac{6}{7}$	$26\frac{16}{21}$	$26\frac{6}{11}$	$26\frac{140}{429}$	24

* Probability for $2n = \infty$, $k > 0$, is given by

$$\frac{9k^3 + 9k^2 + 168k + 208}{216 \cdot 2^k}.$$

is quite good agreement between the curves for finite n and formula (3) at the practically significant percentage points. The situation for very small

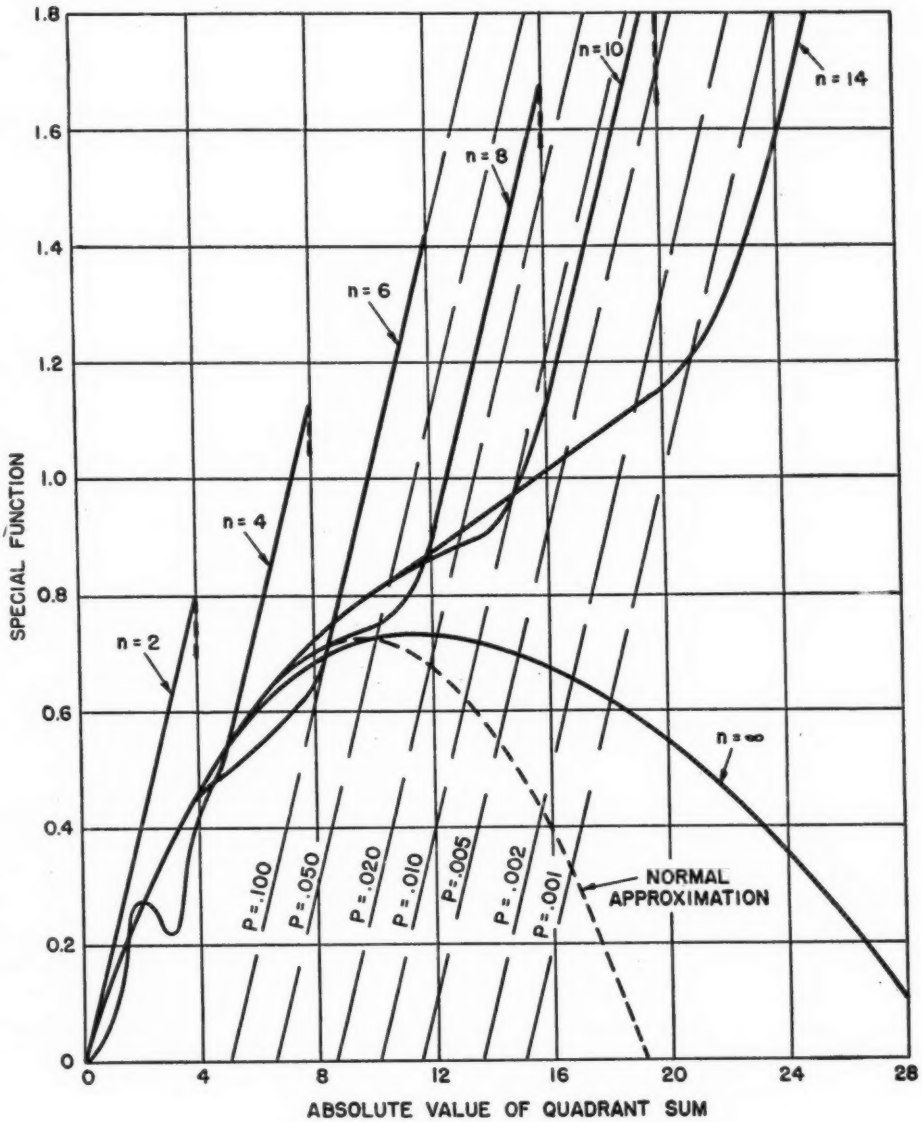


FIG. 3. Comparative relationships for finite and infinite sample sizes and normal approximation to the infinite sample size

probabilities suggests a careful consideration of the limiting behavior of the quadrant sum distribution (see section 10).

The device for samples of $2n + 1$ deserves a word of justification. If there is no association, the $2n + 1$ y-values are randomly paired with the $2n + 1$ x-values, and, in particular, the y-value paired with the x-median is randomly selected. If we pair it with the (randomly selected) x-value which was paired with the y-median we still have random pairing. The pairing of the $2n$ pairs is random, although neither the x-values nor the y-values make up a sample. The randomness of pairing is all that has been used in the discussions of this section.

7. Extension to higher dimensions. The same ideas that underlie the quadrant sum test for two variables may be extended *in several ways* to give tests for various types of association among three or more variables. Only one three-variable case will be discussed here, leaving further extension to the reader.

Given three variables, x, y, and z, and a sample of matched observations on these, it is clearly possible to use the simple quadrant sum test for two variables to investigate association between x and y separately, between y and z separately, and between z and x separately. If the Pearson coefficient of correlation were being computed and were found to be close to zero for each of these pairs, it would be assumed that there was no detectable association through the second moments. In a trivariate normal or Gaussian distribution, where the first and second moments determine the whole distribution, if there is independence between the separate pairs of variables, there is no possibility of a three-way association. It is of some interest, however, to notice that a corner sum test can be devised that will measure the effect of such triple association in case it does exist.

Consider the octants into which the three median planes for x, y, and z, respectively, divide the three dimensional scatter diagram and label the octants alternately plus and minus, in the manner suggested by Fig. 4. More precisely, an octant is counted as plus if an odd number, that is three or one, of the variables are greater than the medians of the sample, and the remaining octants are labelled minus. It is clear that we may repeat the process of coming in along each axis passing from observation to observation as long as they remain in a region of fixed sign, and writing down as a contribution to the final or octant sum the number of such consecutive elements and the sign of the region in which they were found. There will be six terms rather than four, as was the case for the test based on quadrants, and so a new set of significance levels will be required. Table 3, following, lists the situation for a very large sample.

The situation has been sketched for the case of $2n$ triples. If there are $2n + 1$ triples, then we may have trouble with the medians again. However, a similar device works, except that we must agree on a last variable in order to form the synthetic triples uniquely. For example, consider the triples $(m, 3, 5)$, $(9, m, 1)$, $(12, 4, m)$, where m denotes the median. Taking the order in which the variables are written, we get $(12, 3, 5)$ and $(9, 4, 1)$ as the synthetic triples. Other

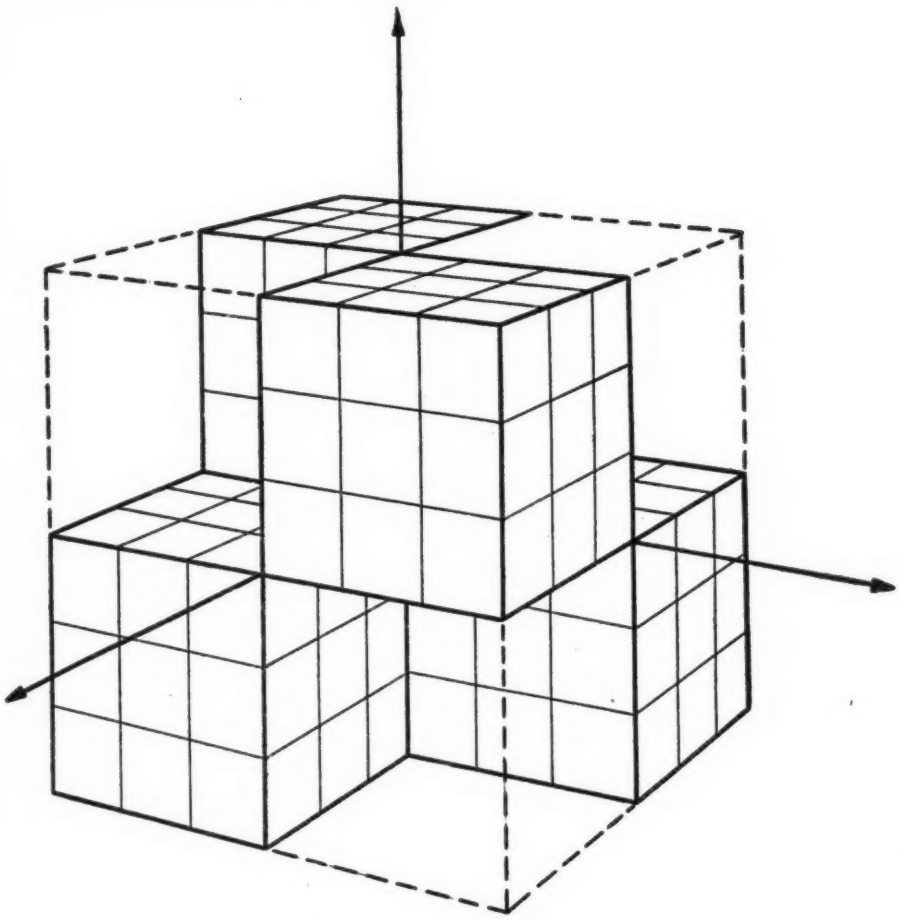


FIG. 4. Octant schematic—solid sections taken as positive

TABLE 3

Working significance levels for the magnitudes of the octant sum

Significance Level	Magnitude of Octant Sum*
10%	11
5%	13
2%	15
1%	16
0.5%	18
0.2%	20
0.1%	21

* Computed for large samples only and based on normal approximation, see section 11 for discussion of this and higher dimensional cases.

orders would yield (9, 3, 5) and (12, 4, 1) or (9, 3, 1) and (12, 4, 5). This slight dissymmetry is not pleasing but should give no difficulty.

8. Nongraphical example. The following example of 78 successive observations of four variables shows how this test may be applied without plotting and how simple the computation still remains. The data concern a metallurgical

TABLE 4

Excerpt from Tippett's Table

Time T*	Fuel F*	Material M*	Articles A*	Duration D*
1 −	246 +	1457 −	1895 +	168.5 +
2 −	196 −	2078 +	2121 +	152 +
3 −	192 −	1278 −	1437 −	153 +
4 −	202 +	1398 −	1497 −	145 −
5 −	206 +	1944 +	1592 +	153 +
6 −	218 +	1464 −	1506 −	147.5 −
7 −	155 −	1541 +	1762 +	152 +
8 −	201 +	1502 +	1818 +	144.5 −
9 −	211 +	1950 +	1144 −	151.5 +
10 −	236 +	1768 +	1654 +	151.5 +
etc. to				
78 +	185 −	1536 +	1442 −	152 +
Median	Median	Median	Median	Median
39.5	199	1474	1588	149.5

* Location of observation relative to column median; + = above; − = below.

Tippett's correlations (based on lightly rounded data)

$$r_{FM} = +0.243$$
$$r_{FA} = +0.266$$
$$r_{MA} = +0.681$$
$$r_{FM.A} = +0.088$$
$$r_{FMA.} = +0.141.$$

problem in mass production and are taken from L. H. C. Tippett, Table XXII, page 63 [2]. An excerpt from the data is given in Table 4 together with Tippett's calculated correlations. This table also shows the preliminary marking of each individual measurement as above (+) for its variable, below (−), or on the median (0). From this table we see, for example, that increasing T contributes a term −3 to the quadrant sum for T and D. It is often desirable to prepare auxiliary tables to assist in computing the components of the quadrant

and hyperquadrant sums. Such a table is Table 5 for low values of Fuel ($F-$) arranged in consecutive ascending numerical order. The entries on this table for the five columns headed F, T, M, A, and D are directly comparable to the entries in Table 4. For example, $F = 155$ is $-$ with respect to the fuel median and $T = 7, -$; $M = 1541, +$; A $= 1762, +$; $D = 152, +$. The double, triple, quadruple and quintuple headed columns contain simply the algebraic multiplication of the signs in the appropriate T, M, A, or D columns. Thus, TM for $F = 155$ is $-$, MAD is $+$, and $TMAD$ is $-$. The contribution to each quadrant or hyperquadrant sum is simply the count of the consecutive like signs from the top of a column. For column AD, we have 7 consecutive $+$ signs and since the contribution is to FAD and F is $-$, the contribution in this case to the octant sum is -7. The results from the ten tables of which Table 5

TABLE 5

Sample Table for One Component of Quadrant and Hyperquadrant Sums. Low Values of Fuel ($F-$)

Fuel F	T	M	A	D	TM	TA	TD	MA	MD	AD	TMA	TMD	TAD	MAD	$TMAD$
98 −	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
135 −	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
140 −	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
146 −	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+
147 −	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+
149 −	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−
151 −	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−
153 −	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+
155 −	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−

Contributions to Sums

FT	FM	FA	FD	FTM	FTA	FTD	FMA	FMD	FAD	$FTMA$	$FTMD$	$FTAD$	$FMAD$	$FTMAD$
−2	+4	+7	+8	+2	+2	+2	−4	−4	−7	−2	−2	−2	+4	+2

is a sample are then carried to the summary computation shown in Table 6. The contribution from Table 5 is shown on line F−. The totals are computed and their probabilities of occurrence determined.

9. Serial example. The following example, a sample of 144 observations of the thickness of inlay for relay springs cut consecutively from a single sheet of material, allows us to compare the resolution of the present test with that of the serial product-moment correlation. The data are from Shewhart [1, 1941, Table 1] and the serial correlations from lag 1 to lag 22 are from recent calculations by Miss Dorothy T. Angell. The procedure for calculating the serial quadrant sums is similar to that for obtaining the sums for section 8. A table is prepared to show the observed consecutive order of the numerical values and each is identified as above (+), below (−), or on the median (0). This gives a

TABLE 6

Summary Computation Table for Quadrant and Hyperquadrant Sums

From Table	*TF*	*TM*	*TA*	*TD*	*FM*	*FA*	*FD*	*MA*	*MD*	*AD*	*TFM*	*TFA*	*TFD*	*TMA*	*TMD*	*TAD*	*FMA*	*FMD*	*FAD*	*MAD*	*TFMA*	*TFMD*	*TFAD*	*TMAD*	*FMAD*	*TFMAD*
T +	−6	+1	−2	+3							−1	+2	−3	−1	+1	−2					+1	−1	−1	+2		+1
T −	−1	+1	−2	−3							+2	−1	−1	+1	+1	−2					+3	+2	+1	−1		+4
F +	+4				+4	+1	+1				+5	+1	+1				+1	+1	+3		+1	+1	+3		+3	+3
F −	−2				+4	+7	+8				+2	+2	+2				−4	−4	−7		−2	−2	−2		+4	+2
M +		+1			+2			+5	+13		+1			+1	+1		+2	+2		+5	+1	+1		+1	+2	+1
M −		−2			+6			+25	+3		+2			+2	+2		−6	−3		−3	−2	−2		−2	+3	+2
A +			−2			−1		+3		+7		+1		−2		−2	−1		−1	+3	+1		+1	−2	−1	+1
A −			−1			+5		+1		+4		+1		+3		+1	−1		−4	−1	−3		−1	−3	+1	+3
D +				+1			+2		+1	+2			+1		+3	+1		+1	+2	−1		−2	−1	−2	+1	+2
D −				−1			+3		+1	+7			+1		+2	+1		−1	−3	+1		+2	+1	+2	+1	+3
Totals Quadrant Sums.	−5	+1	−7	0	+16	+12	+14	+34	+18	+20																
Octant Sums.....											+11	+6	+1	+4	+10	−3	−9	−4	−10	+4						
Hexadecant Sums.....																					0	−1	+1	−5	+14	
Dotriacontant Sums...																										+22
Probability (%) ≤.....	36	92	19	100	0.2	2	1	0.1	0.1	0.1	9	37	94	57	12	68	16	57	12	57	100	95	95	52	6	0.6
Significant at 5%.....					*	*	*	*	*	*																*
Significant at 1%.....					*		*	*	*	*																*
Significant at 0.2%....					*			*	*	*																

table similar to one of the elements, say Fuel, in Table 4. Four computation tables similar to Table 5 are required, one for the equivalent of moving from the right, one from below, one from the left, and one from the top of a lag correlation scatter diagram. One table from each direction will take care of all lags. In the first, the marginal entries are the observed values listed in descending numerical order. Opposite these are recorded from the previous table the signs associated with observations for each lag with respect to each entry. The second table would record the signs relating to the lags from the observed values arranged in ascending order. The third table would record the signs relating to leads from the observed values arranged in ascending order and the fourth, the signs relating to leads from the observed values arranged in descending order. The sign of the contribution from each group is the algebraic product of the sign of the run and the sign of the marginal entries. The length of run is determined in the same way as in Table 5. Table 7 illustrates the procedure

TABLE 7

Relation of Lagged Observations to Median (+ = above, − = below) for Smallest Observations in Ascending Order

Thickness	Lag																								
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
2	−	−	+	−	−	+	−	+	−	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−
3	−	−	+	−	+	−	−	−	−	+	−	−	+	−	+	−	+	−	+	+	+	+	+	+	+	+
8	−	−	−	+	−	−	−	−	+	+	−	+	+	−	+	+	−	−	+	+	−	−	−	−	−	−
10	−	+	−	−	−	−	+	+	−	+	+	−	+	+	−	−	+	+	−	−	−	−	−	−	+	−
13	−	−	−	−	−	+	+	+	−	+	+	+	+	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	( )	(+)	(+)	(+)	(+)
17	−	−	−	+	−	−	+	−	+	−	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	−
17	−	+	−	+	+	+	+	+	−	−	+	+	−	−	−	−	−	−	−	+	−	-	+	−	+	+
18	−	−	−	−	+	+	−	+	+	−	+	+	−	−	+	+	−	−	−	−	−	−	+	−	+	−
•		+3	−2	+2	+1	−1	+3	−1	+2	−5	+3	−1	−7	−1	−3	−1	−2	−1	−3	−3	−2	−2	−2	−2	+1	+1

• Contribution to Serial Quadrant Sum.

of determining the contribution from lags associated with the observations arranged in ascending order.

Two serial quadrant sums may be computed—a circular serial quadrant sum or a noncircular serial quadrant sum. Circular items arise from considering that the beginning of the set of observations is a continuation of the end in the same way that this assumption is made in computing circular serial correlation coefficients. In Table 7, circular items are shown in parentheses and are omitted in calculating noncircular sums. In the particular table shown, the count of the run lengths was identical for both types of sum, but in other cases this may not be the case. Since the serial quadrant sum is relatively insensitive to sample size, the noncircular serial quadrant sum has for all practical purposes the same distribution as the circular quadrant sum. The correspondence in this case between the serial correlation coefficient for each lag up to 22 and the respective values of the two types of serial quadrant sums is shown in Fig. 5.

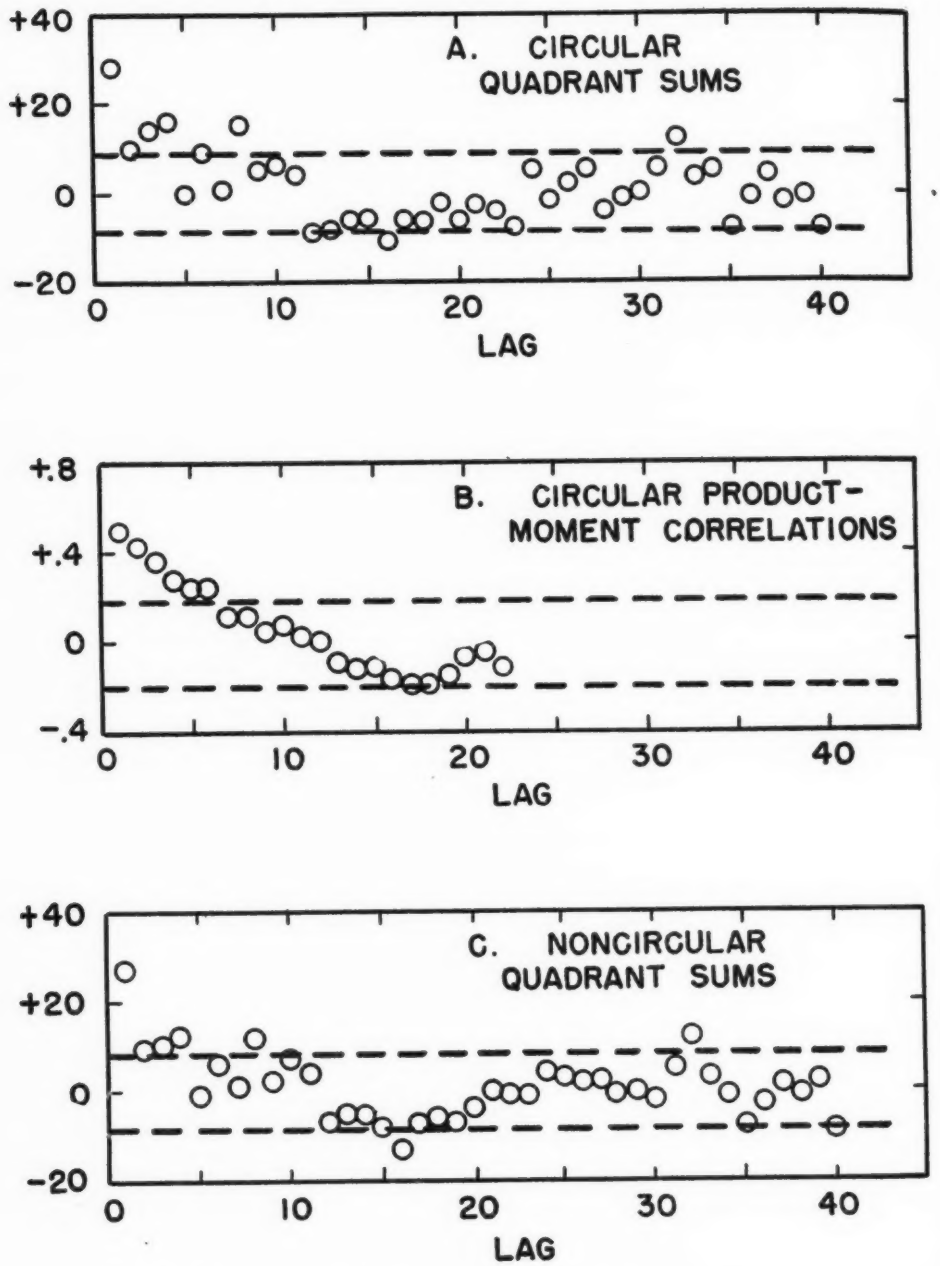


FIG. 5. Comparative performance on a serial (autocorrelative) example

10. Convergence to the limiting distribution. We shall consider several chance sums. One of these is S, which has the limiting distribution discussed

in section 6. Another is S_k', which is the sum of four independent terms, each distributed according to the limiting distribution curtailed at $\pm k$. Its generating function is

$$G_k(x) = \left(\sum_{i=1}^{k} 2^{-(i+1)} x^i + \sum_{i=1}^{k} 2^{-(i+1)} x^{-i}\right)^4.$$

The total probability assigned to $S_k' = -k, -(k-1), \cdots, k$, is less than unity, so that there is nonzero probability that S_k' is not defined. The third is S_n, the quadrant sum itself, whose generating function is (5), and the fourth is the result of the same sort of curtailment applied to S_n. It will be denoted by $S_{n,k}$ and its generating function is

$$G_{n,k}(x) = \sum_j \frac{(j!)^2((n-j)!)^2}{(2n)!}\left(\sum_{i=1}^{k}\binom{n-i-1}{n-j-1}x^i + \sum_{i=1}^{k}\binom{n-i-1}{j-1}x^{-i}\right)^4.$$

This again corresponds to a total probability less than unity.

It is clear that

$$\Pr(S_{n,k} = m) \leq \Pr(S_n = m)$$

and

$$\Pr(S_k' = m) \leq \Pr(S = m).$$

We shall soon show that

$$\lim_{n\to\infty} \Pr(S_{n,k} = m) = \Pr(S_k' = m) \tag{6}$$

and this will imply that

$$\lim_{n\to\infty} \Pr(S_n = m) = \Pr(S = m)$$

which is the desired result. The implication runs as follows: given ϵ, we can choose k so large that

$$\Pr(S_k' \text{ defined}] \geq 1 - \epsilon/3$$

whence

$$|\Pr(S_k' = m) - \Pr(S = m)| \leq \epsilon/3$$

and then choose n so large that

$$|\Pr(S_{n,k} = m) - \Pr(S_k = m)| \leq \epsilon/(24k + 6)$$

$$\text{for } m = -4k, -4k+1, \cdots, 4k$$

whence

$$\Pr(S_{n,k} \text{ defined}) \geq 1 - \epsilon/3 - \frac{8k+1}{24k+6}\epsilon \leq 1 - \frac{16k+3}{24k+6}\epsilon$$

and hence

$$| \Pr(S_{n,k} = m) - \Pr(S_n = m) | \leq \frac{16k + 2}{24k + 6} \epsilon$$

this inequality holding automatically for $| m | > 4k$. Hence,

$$| \Pr(S_n = m) - \Pr(S = m) |$$
$$\leq | \Pr(S_n = m) - \Pr(S_{n,k} = m) | + | \Pr(S_{n,k} = m) - \Pr(S'_k = m) |$$
$$+ | \Pr(S'_k = m) - \Pr(S = m) | \leq \frac{16k + 2}{24k + 6} \epsilon + \frac{1}{24k + 6} \epsilon + \tfrac{1}{3}\epsilon < \epsilon$$

This method is clearly of general application in such problems.

We turn now to the proof of (6). The expression for $G_{n,k}(x)$ shows that we may consider it the result of the following process: the integer j is a chance quantity with the distribution

$$\Pr(j = j_0) = \frac{(n!)^2}{(2n)!} \binom{n}{j_0}^2.$$

For fixed j, $G_{n,k}$ is the average over j of

$$G_{n,k,j}(x) = \left[\sum_{i=1}^{k} \frac{\binom{n-i-1}{n-j-1}}{\binom{n}{j}} x^i + \sum_{i=1}^{k} \frac{\binom{n-i-1}{j-1}}{\binom{n}{j}} x^{-i} \right]^4.$$

The first of these relations shows that j/n converges stochastically to $\frac{1}{2}$ as n approaches infinity. The second shows, since

$$\frac{\binom{n-i-1}{n-j-1}}{\binom{n}{j}} = \frac{(n-i-1)!(n-j)!j!}{(n-j-1)!(j-i)!n!} = \frac{(n-j)(j)(j-1)\cdots(j-i+1)}{n(n-1)(n-2)\cdots(n-i)}$$

$$\frac{\binom{n-i-1}{j-1}}{\binom{n}{j}} = \frac{(n-i-1)!(n-j)!j!}{(n-j-i)!(j-1)!n!}$$

$$= \frac{(n-j)(n-j-1)\cdots(n-j-i+1)j}{n(n-1)\cdots(n-i)}$$

and both of these converge stochastically to $2^{-(i+1)}$ as n approaches infinity, that $G_{n,k,j}(x)$ converges stochastically to $G_k(x)$. Since these curtailed generating functions involve only powers of x in the finite range between $-4k$ and $+4k$, the limiting relation (6) follows at once.

11. Effectiveness of normal approximation. Fig. 3 shows the relation between the asymptotic distribution of the quadrant sum for large n and a normal

distribution with variance 24, i.e., the same variance as that of the asymptotic distribution. The normal approximation is calculated from

$$\Pr(|S_n| \geq m) \approx \Pr\left(x \geq \frac{m - \frac{1}{2}}{\sqrt{24}}\right)$$

where x is normally distributed with zero mean and unit variance. The asymptotic and normal curves agree surprisingly well out to the 5% point, and an error of a full unit in the significance level first occurs beyond the 0.5% point.

Since the asymptotic distributions for the quadrant, octant, hexadecant, dotriacontant,—, sums become more and more normal, the normal approximation will be even better for higher dimensions. In r dimensions, this approximation consists in treating

$$\frac{|S_n| - \frac{1}{2}}{\sqrt{12r}}$$

as the absolute value of a standard deviate. This should be quite adequate for large samples and $r \geq 4$.

12. Unsolved problems. The central unsolved problem in connection with the quadrant sum is:

(1) What is the operating characteristic?

This has as a corollary the more general question:

(2) How can the operating characteristic of a nonparametric test be described so as to be useful to the users of the test?

There are, of course, minor problems which are much more easily soluble. A few, listed in order of practical importance, are:

(3) What is the effect on the significance levels of the use of lagged values of x as values of y?

(4) What are the exact distributions for moderate n in three or more dimensions?

(5) Do the analogous limiting distributions hold for three or more dimensions?

(6) What is a better approximation to the limiting distribution for moderate n?

To encourage others to solve some of these, we close with the assurance that they have our good wishes.

REFERENCES

[1] W. A. Shewhart, "Contribution of statistics to the science of engineering," *University of Pennsylvania Bicentennial Conference*, Volume on Fluid Mechanics and Statistical Methods in Engineering, University of Penn. Press, 1941, pp. 97–124. Also Monograph B-1319, *Bell Tel. System Tech. Pub.*, 1941.

[2] L. H. C. Tippett, *Statistical Methods in Industry*, pamphlet published by Iron and Steel Research Council of the British Iron and Steel Federation, 1943.

[3] A. Wald, "An extension of Wilks' method for setting tolerance limits," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 45–55.

DISCRIMINANT FUNCTIONS

By George W. Brown

Iowa State College

1. Introduction: In the following sections the development of discriminant function techniques is approached from an elementary point of view, considering first an essentially trivial problem, then working up to the more complex situations which may be handled by discriminant function methods. No attempt has been made to follow the pattern of the historical development in this process, and no consistent attempt has been made to allocate proper credit, in the text, to those individuals responsible for the introduction and exploitation of these methods. A more or less exhaustive bibliography of discriminant function applications and related theory is given at the end of this paper.

Some historical perspective may be gained, however, from a very sketchy consideration of the early background of the subject. The first published application of the discriminant function seems to have been the work of Barnard (1935 [1]) on craniometry, following the suggestion of R. A. Fisher. Meanwhile P. C. Mahalanobis (1927, [30]; 1930, [31]) and, in this country, Hotelling (1931, [25]) had been concerned with a closely related problem, the construction of measures of the "distance" between two sets of multiple measurements, for which Karl Pearson's (1926, [34]) coefficient of racial likeness was not wholly adequate. Fisher (1936, [18]) gave a further example of the method and showed (1938, [19]) the relation between his work and that of Hotelling (1931, [25]; 1936, [27]). Thus the theory of discriminant function analysis proper is about ten years old, but is intimately related to researches which go back a few more years.

A simple problem: Consider the very simple case of a single measurement, say ξ, which may be made in each of two populations, and let us suppose, for the sake of discussion, that ξ is normally distributed, with unit variance, in each population, but with possibly different means in the two populations.

Let

$$E_1(\xi) = \alpha - \beta$$

$$E_2(\xi) = \alpha + \beta$$

be the mean values of ξ over the two populations, with $\beta > 0$. As an example, we may consider the pH measurements of Iowa soil samples (Cox and Martin, [12]), for two soil populations, distinguished by the presence or absence of Azotobacter. From 100 samples containing Azotobacter and 186 samples containing no Azotobacter, we have the estimated averages of pH equal to 7.423 and 6.015 respectively, with an estimated standard error of .625 within populations (see Fig. 1).

$$\hat{\alpha} = 6.719$$

$$\hat{\beta} = .704$$

$$\hat{\sigma} = .625$$

$$\hat{\beta}/\hat{\sigma} = 1.13.$$

Let us suppose further that ξ is the only measurement available on a single individual, not knowing to which of populations 1 and 2 the individual belongs.

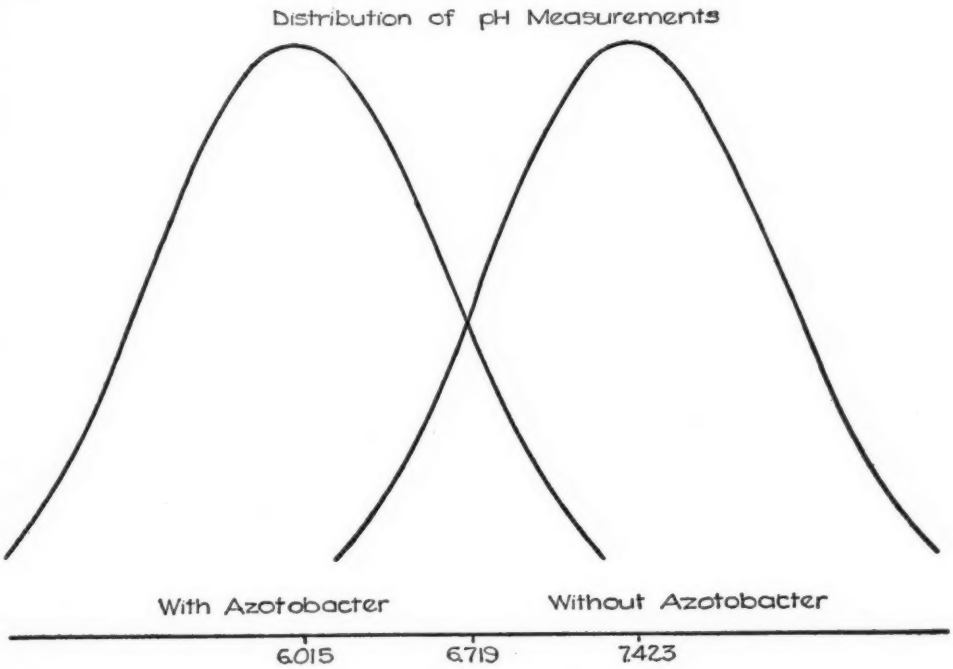


FIG. 1

The problem is to classify this individual as a member of population 1 or population 2. It is clear that ξ furnishes the only information on which to base a decision, and that essentially the only procedure available is to choose a number, say ξ_0, such that we choose population 1 when $\xi < \xi_0$ and population 2 when $\xi > \xi_0$. Furthermore, it is evident that the expected accuracy of classification depends on the size of β. If we wish to have equal risks of misclassification for members of the two populations we choose $\xi_0 = \alpha$. Then the probability of misclassification is given by $P\{\epsilon > \beta\}$, where ϵ is a normal deviate with unit variance. As one would expect, the probability of misclassification tends to 0 as $\beta \to \infty$ and tends to $\frac{1}{2}$ as $\beta \to 0$. In the Azotobacter example, if we assume that the estimates given are the population values, we choose $\xi_0 = 6.719$. The

ratio $\hat{\beta}/\hat{\sigma} = 1.13$ is exceeded approximately 13% of the time in sampling from the normal distribution, leading to .13 as the probability of misclassification.

Consider now the slightly more general situation in which we consider a fixed variate, say w with measurements ξ distributed, for fixed w, with a mean of the form $\alpha + \beta w$. This is the standard regression situation. As before assume that ξ is normally distributed about this mean with unit variance, that is

$$\xi = \alpha + \beta w + \epsilon$$

where α and β are constants, w may take on any or all real values, and ϵ is a normal deviate. Note that if w is restricted to take on only two values the structure reduces to the first structure considered. An example of the continuous type might be constructed by considering w as genotypic yield of grain and ξ a phenotypic measure of yield (Smith, [36]).

The simple problem formulated for the two-population case may be reformulated here as follows: Given the relationship $\xi = \alpha + \beta w + \epsilon$, and given ξ for an individual for which no other information is known, how shall we estimate w? For selective breeding the problem may be to select individuals for which w is at one end of the scale, rather than to estimate w itself. Whatever decision is to be made, it is still clear that ξ furnishes the only available information, and that the certainty of the decision is a function of β. Since $(\xi - \alpha)/\beta = w + \epsilon/\beta$, the variance of this estimate of w is $1/\beta^2$. Note that confidence intervals for w, given ξ, may be constructed from the normally distributed quantity $\xi - \alpha - \beta w$.

It should be pointed out that in the usual regression case we are interested in predicting ξ for given w, with the hypothesis as stated above, whereas in this case ξ will be observed, and the problem is that of estimating, as a parameter of the distribution of ξ, the fixed variate w.

Obviously β must not vanish if ξ is to perform any discrimination among w values. In practice, of course, α and β will not be given as known values and the variance of ϵ will not be known, but a finite set of observations may be available, for which w values are known and ξ has been observed. The usual analysis of variance provides a significance test for the non-vanishing of β, which is equivalent to testing for the significance of the regression of ξ on w.

It is to be noted that this analysis reduces to the conventional between-within analysis (F or t-test) when we have the special case of two populations. Moreover, if we had treated ξ as the fixed variate instead of w, and considered the regression of w on ξ, the Analysis of Variance would have differed only in replacing $\Sigma(\xi - \bar{\xi})^2$ throughout by $\Sigma(w - \bar{w})^2$ and the relevant F-test would have been unchanged.

When probabilities of misclassification are estimated from finite samples, as in the soil classification example, there are three sources of error, sampling error in the estimate of the separation value ξ_0, sampling error in the estimate of the distance between the population means, and sampling error in the estimated standard deviation of ξ within populations. It does not appear difficult to set up confidence intervals for the probability of misclassification, assuming repeated classification of individuals given fixed initial samples.

2. The one-dimensional discriminant function. We have been dealing so far with the simple situation in which only one measurement per individual is available for purposes of discrimination. Suppose we still have this measurement, call it ξ_1, now, but we have other measurements as well, say $\xi_2, \cdots, \xi_p$. As before $\xi_1 = \alpha_1 + \beta w + \epsilon_1$. For the moment suppose that the remaining measurements have mean values independent of w, so that

$$\xi_m = \alpha_m + \epsilon_m, \qquad (m = 2, \cdots, p),$$

and let us assume also that the $\{\epsilon_m\}$ are mutually independent, $(m = 1, 2, \cdots, p)$ and are normal deviates with unit variance. It is safe to assume that nobody would ever argue, in this case, that the measurements $\xi_2, \cdots, \xi_p$, provide information about the w value for an individual. If, then, we were so fortunate that we were in this situation, and knew so, we could say that ξ_1 is our discriminant function, since, if any discriminating is to be done, ξ_1 has to do it.

TABLE 1

Analysis of Variance for Regression

	d.f.	Sums of Squares
Regression	1	$r^2\Sigma(\xi - \bar{\xi})^2$
Error	$N - 2$	$(1 - r^2)\Sigma(\xi - \bar{\xi})^2$
Total	$N - 1$	$\Sigma(\xi - \bar{\xi})^2$

$$r = \frac{\Sigma(\xi - \bar{\xi})(w - \bar{w})}{\sqrt{\Sigma(\xi - \bar{\xi})^2 \Sigma(w - \bar{w})^2}}$$

Suppose, now that the measurements $\xi_1, \xi_2, \cdots, \xi_p$ are not explicitly available, but that we are able to observe a linearly equivalent set $x_1, x_2, \cdots, x_p$, related to the $\{\xi_m\}$ by the transformation

$$x_m = \sum_{n=1}^{p} l_{mn} \xi_n$$

where the l_{mn} are unknown. For fixed w, x_m has expected value

$$\sum_{n=1}^{p} l_{mn} \alpha_n + l_{m1} \beta w = a_m + b_m w,$$

so that in general each x_m observation provides information about w. Moreover, the x_m are not in general mutually independent; it is evident that the population matrix of variances and covariances for fixed w is given by $\sigma_{mn} = \sum_{k=1}^{p} l_{mk} l_{nk}$.

As an example of a set of correlated measurements, consider the Azotobacter example referred to above. In addition to pH values, determinations of avail-

able phosphate content and total nitrogen content were made on soil samples in each of the two populations. Means were as follows:

	pH	*Phosphate*	*Nitrogen*
Mean of 100 samples with Azotobacter	7.423	133.120	29.400
Mean of 186 samples without "	6.015	51.113	21.140
Mean difference	1.408	82.007	8.260

Clearly the differences are proportional to the hypothetical b_m's. The variance-covariance matrix, estimated from the 284 degrees of freedom within populations, is given by Table 2.

TABLE 2

	pH	Phosphate	Nitrogen
pH	111.0879	2,292.7192	198.4026
$284(\sigma_{mn})$ = Phosphate		1,042,799.1890	5,066.2645
Nitrogen			29,422.3655

Estimated correlation coefficients within populations are not large, .213 for *p*H and Phosphate, .110 for *p*H and Nitrogen, and .029 for Phosphate and Nitrogen.

Another example is furnished by Fisher's Iris measurements [8], providing sepal length, sepal width, petal length, and petal width for each of 50 individuals of Iris setosa and 50 individuals of Iris versicolor. This example is an unfortunate one in that either petal length or petal width alone is sufficient to discriminate the two populations as completely as anybody has a right to expect anytime. The petal lengths, for example, vary between 1.0 and 1.9 cm. for the 50 setosa, and between 3.0 and 5.1 cm. for the 50 versicolor.

Let us proceed, under the assumption that available measurements, x_m, are distributed normally about mean values $a_m + b_m w$, with variance covariance matrix σ_{mn} for fixed w, keeping in mind the underlying model of $\xi_1, \xi_2, \cdots, \xi_p$, with

$$x_m = \sum_{n=1}^{p} l_{mn}\xi_n, \qquad \xi_1 = \alpha_1 + \beta w + \epsilon_1; \qquad \xi_2 = \alpha_2 + \epsilon_2; \cdots; \xi_p = \alpha_p + \epsilon_p.$$

The skeptic may wish to grant the first part of our assumptions without granting the hypothetical structure of ξ's underlying the x's. Hotelling's work [27] shows that such an underlying structure of ξ's may always be provided, given the distribution of x's for fixed w. In other words, a distribution of x's for fixed w leads essentially uniquely to an underlying ξ model.

The discriminant function, given σ_{mn}, a_m and b_m, for $m, n, = 1, 2, \cdots, p$, is

$$X = \sum_{m,n=1}^{p} \sigma^{mn} b_m x_n = \sum_{n=1}^{p} l_n x_n$$

where

$$l_n = \sum_{m=1}^{p} \sigma^{mn} b_m, \quad \text{and} \quad \sigma^{mn}$$

is the reciprocal matrix to σ_{mn}. That is σ^{mn} are the solutions of the linear systems [17]

$$\sum_{s=1}^{p} \sigma^{ms}\sigma_{sn} = 0 \quad \text{if} \quad m = n; \qquad m, n, = 1, 2, \cdots, p$$

$$\sum_{s=1}^{p} \sigma^{ms}\sigma_{sm} = 1; \qquad m = 1, \cdots, p.$$

That X, as defined above, is properly called the discriminant function will become evident immediately. Putting $b_m = l_{m1}\beta$, $x_n = \sum_{k=1}^{p} l_{nk}\xi_k$, we have

$$X = \beta \sum_{m,n,k,} \sigma^{mn} l_{m1} l_{nk} \xi_k .$$

Recalling that the σ^{mn} are reciprocal to $\sigma_{mn} = \sum_{k} l_{mk}l_{nk}$, it can be seen that $\sum_{mn} \sigma^{mn}l_{m1}l_{nk} = 1$ if $k = 1$, and vanishes for $k = 1$. It follows that

$$X = \beta\, \xi_1 ,$$

in other words, X calculated as $\sum_{mn} \sigma^{mn}b_m x_n$ from known population quantities is proportional to the hypothetical ξ_1, the only one of the underlying measurements which is related to w, thus justifying the term discriminant function for X. It is clear that any other linear function of the x's is also a linear function of the ξ's, and can discriminate, at best, only as well as X itself, since all the ξ's are independent of w, with the exception of ξ_1. X itself discriminates w to the same extent that ξ_1, were it available, would discriminate.

The degree of discrimination of w's depends, as indicated in the previous section, on the ratio of the mean square of ξ_1, among w's (mean square for regression), to the mean square of ξ_1 for fixed w (mean square for error). Since X is proportional to ξ_1, the same is true when X is substituted for ξ_1. It turns out, of course, that X is that linear combination of x's for which the ratio of the mean square for regression to the mean square for error is a maximum, or, what is the same thing, X is that linear combination of X's which has the maximum correlation with w. From any point of view X appears to be the logical function of x's to compute. It is clear that λX is precisely as good as X, if λ is any constant.

In the two population case, where w takes on only two values, X is evidently proportional to $\Sigma\sigma^{mn}(\mu_{m1} - \mu_{m2})x_n$, where μ_{m1} and μ_{m2} are the mean values of x_m in the two populations. X is here the particular linear combination of x's for which the ratio of the mean square between populations to the mean square within populations is a maximum. The value of this ratio, which measures the degree of discrimination possible, depends on the spread of the means of X between the populations, or in general, on the spread of the means of X over some given distribution of w's. Given σ_{mn} and b_m the larger the spread of w values the better overall discrimination will be obtainable. On the other hand, the coefficients for X depend only on σ_{mn} and b_m.

Since X is proportional to ξ_1, it follows that the discriminant function is invariant under non-singular linear transformation of the x's, that is, if some set of y's, linearly dependent on the x's, had been observed, together with their means, variances and covariances, the discriminant values would not have changed. This invariance is obviously a desirable property, and as such was one of the goals of Fisher, Hotelling, and Mahalanobis. One more property of the discriminant function is of interest; X is essentially equivalent to the maximum likelihood estimate of w.

In our statistical model w plays the role of a fixed variate or population parameter, and the x's have a joint distribution about linear functions of w as means. Suppose now that (σ_{mn}) and $\{b_m\}$ are estimated from an analysis of variance and covariance on data for which w as well as x values are known. The problem of estimating w for a single individual whose x measurements are given resolves into a two-stage estimation process, the first stage being the estimation of (σ_{mn}) and $\{b_m\}$ from the initial data, the second stage being the estimation of w by the discriminant function whose coefficients are computed from the estimated (σ_{mn}) and $\{b_m\}$. It has already been pointed out that X is the linear combination of x's which has greatest correlation with w. It turns out, then, that the coefficients of X are proportional to those which would have been obtained from a formal regression analysis of w on $x_1, x_2, \cdots, x_p$, considering the x's as independent variables and w as dependent variable, a direct interchange of roles as compared with the statistical model we have assumed. Of course two linear functions differing only by a factor of proportionality are equivalent in discrimination. If the formal analysis of variance is carried out for testing the significance of the regression of w on $x_1, x_2, \cdots, x_p$, the relevant F ratio remains a valid test for the non-vanishing of the b_m in spite of the inversion of dependent and independent variables. The analysis of variance is given in Table 3.

R is, of course, the conventional multiple correlation coefficient. An equivalent analysis can be carried out for X itself, allowing sufficient degrees of freedom for the estimation of the constants in X, as given in Table 4.

This analysis is proportional to the analysis given above. It might be noted that the mean square corresponding to error sum of squares in this analysis is $\Sigma\sigma^{mn}b_m b_n$, which is X evaluated for $x_n = b_n$, $(n = 1, 2, \cdots, p)$.

In the Azotobacter example, Cox and Martin arrive at a discriminant function which has the analysis given in Table 5.

It is evident that the difference between populations is highly significant. The choice of scale for X in this case forces the sum of squares within populations to be equal to the difference between the mean X values for the two populations. Thus the mean X differs by .021777 for the two populations, and has an esti-

TABLE 3

Analysis of Variance for Regression

	d.f.	Sums of Squares
Regression	p	$R^2\Sigma(w - \bar{w})^2$
Error	$N - p - 1$	$(1 - R^2)\Sigma(w - \bar{w})^2$
Total	$N - 1$	$\Sigma(w - \bar{w})^2$

TABLE 4

Analysis of Variance for X on w

	d.f.	Sums of Squares
Regression	p	$R^2\Sigma(X - \bar{X})^2$
Error	$N - p - 1$	$(1 - R^2)\Sigma(X - \bar{X})^2$
Total	$N - 1$	$\Sigma(X - \bar{X})^2$

TABLE 5

Analysis of Variance of Discriminant Function

	d.f.	Sums of Squares	Mean Square
Between populations	3	.030842	.01028
Within populations	282	.021777	.00007722
Total	285		

mated standard error, within populations, equal to $\sqrt{.00007722} = .008788$. Half the difference, divided by the standard error is the normal deviate corresponding to misclassification, if equal risks are taken. In this case the value of the normal deviate is 1.24, approximately, leading to an estimated probability of misclassification of about .11, which is not very much better than the .13 which one would have obtained if pH alone had been used.

In this problem, as in conventional regression analysis, it is tempting, for

various reasons, to consider the possibility of using smaller sets of classifying measurements. Moreover, a significance test for this situation is in general more interesting, as a practical matter, than the significance test for differences among populations, since the initial presumption is that we are interested in being able to discriminate, on the basis of $x_1, x_2, \cdots, x_p$. Suppose, for example, we wish to test whether the discriminant function $X_{(p)}$ based on $x_1, x_2, \cdots, x_p$ is significantly better than the discriminant function $X_{(r)}$ based on $x_1, \cdots, x_r$, with $r < p$. The relevant test is precisely the same as the test

TABLE 6

Analysis of Variance for Rejecting $x_{r+1}, \cdots, x_p$

		Sums of Squares	d.f.
S_r^2	Regression on	$x_1, \cdots, x_r$	r
S_p^2	Regression on	$x_1, \cdots, x_r, x_{r+1}, \cdots, x_p$	p
$S_p^2 - S^2$	Difference		$p - r$
$S_T^2 - S_p^2$	Error		$N - p - 1$
S_T^2	Total		$N - 1$

TABLE 7

Analysis of Variance for $X = X_0$

	Sums of Squares	d.f.
S_p^2	Regression on X_0	1
S_p^2	Regression on $x_1, \cdots, x_p$	p
$S_p^2 - S_1^2$	Difference	$p - 1$
$S_T^2 - S_p^2$	Error	$N - p - 1$
S_T^2	Total	$N - 1$

calculated formally from the regression of w on the sets $x_1, \cdots, x_r$ and $x_1, x_2, \cdots, x_p$, with the analysis of variance given in Table 6.

Similarly, if we wish to test for the significance of a theoretical discriminant function, X_0, with preassigned coefficients, as compared with X_p, we have again the conventional test calculated from the formal analysis of the regression of w on $x_1, x_2, \cdots, x_p$, as given in Table 7.

As shown by Fisher [21] the relevant F-Test for this hypothesis is computable as

$$F_{p-1,n-p+1} = \frac{n - p + 1}{p - 1} \frac{R'^2}{1 - R'^2}$$

where $R'^2 = R^2(1 - r^2)$, r is the correlation between X and X_0 for fixed w, and R is the multiple correlation for w on $x_1, \cdots, x_p$, or, what is the same thing, the correlation of w and X.

The example of Smith [36] is an example in which the relationships of x's to w have to be estimated from analysis of variance and covariance of data in which the w's are not really known, being related to genotypes. The regression of x's on w is estimated by a generalization of the components-of-variance method, from variance-covariance analyses in which the usual null hypotheses are significantly contradicted. The net effect is that the usual significance tests now fail to hold, although the algebraic calculations are formally equivalent to those given above, once the population relations of x's to w are established. When work of this kind is based on small samples, there is some difficulty in estimating the reliability of the results.

3. Multi-dimensional discriminant functions. Instead of trying to discriminate between two populations or estimate a single parameter w, our problem may be to discriminate among several populations, not necessarily linearly related, or to estimate many independent parameters $w_1, w_2, \cdots, w_s$. Just as a single parameter w is sufficient to distinguish between means of measurements for two different populations, s parameters are sufficient to distinguish between means of $s + 1$ different populations, and exactly s parameters will be required, if no linear relation obtains among the $s + 1$ populations. For example, with three populations, any measurement mean may be given the three possible values α, $\alpha + \beta$, $\alpha + \gamma$, corresponding to $w_1 = w_2 = 0$ for population 1, $w_1 = 1$, $w_2 = 0$ for population 2, and $w_1 = 0$, $w_2 = 1$ for population 3. Geometrically we have to consider a set of parameter values as a point in an s-dimensional space.

The one-dimensional discriminant function admits two very different generalizations in higher dimensions. The practical solution to a particular problem for which s is moderately large may involve a mixture of both generalizations.

Let us generalize our statistical model before discussing the discrimination problem. To avoid complication of algebraic notation, let us for the moment assume $s = 2$. We will now postulate a set of hypothetical measurements $\xi_1, \xi_2, \cdots, \xi_p$, with

$$\xi_1 = \alpha_1 + \beta_1 u + \gamma_1 v + \epsilon_1$$

$$\xi_2 = \alpha_2 + \beta_2 u + \gamma_2 v + \epsilon_2$$

$$\xi_3 = \alpha_3 + \epsilon_3$$

$$\cdot$$

$$\cdot$$

$$\cdot$$

$$\xi_p = \alpha_p + \epsilon_p,$$

where the ϵ_p are independent normal deviates with unit variance, u and v are fixed variates or parameters corresponding to the different populations, and $\alpha_1, \alpha_2, \cdots, \alpha_p, \beta_1, \beta_2, \gamma_1$, and γ_2 are constants. Evidently $\xi_3, \cdots, \xi_p$ can yield no information about u and v; ξ_1 and ξ_2 together contain all the information there is to get about u and v. As before, assume that our data will be in the form of linear combinations $x_m = \Sigma l_{mn}\xi_n$, with unknown coefficients l_{mn}. The variance-covariance matrix within populations, or for fixed u, v, is still given by $\sigma_{mn} = \Sigma l_{mk}l_{nk}$. The mean values of the x's for fixed u, are given by

$$\begin{aligned} E(x_m) &= \Sigma l_{mn}\alpha_n + (l_{m2}\beta_1 + l_{m2}\beta_2)u + (l_{m1}\gamma_1 + l_{m2}\gamma_2)v \\ &= A_m + b_m u + c_m v. \end{aligned}$$

This model is again justifiable on he basis of Hotelling's work.

The first question to ask is whether we can now form two linear combinations of the x's and get rid of $\xi_3, \cdots, \xi_p$ in both, thus providing a two dimensional description of an individual on the basis of $x_1, x_2, \cdots, x_p$. The answer here is in the affirmative, as a result of a direct generalization of the method discussed earlier. If we calculate $X_1 = \Sigma\sigma^{mn}b_m x_n$ and $X_2 = \Sigma\sigma^{mn}c_m x_n$, we are fortunate enough to get

$$\begin{aligned} X_1 &= \beta_1\xi_1 + \beta_2\xi_2 \\ X_2 &= \gamma_1\xi_1 + \gamma_2\xi_2 \end{aligned}$$

with no disturbing elements from $\xi_3, \cdots, \xi_p$. Assuming for now that X_1 and X_2 are not merely proportional, i.e. $\beta_1\gamma_2 - \beta_2\gamma_1 \neq 0$, what do we do with X_1 and X_2?

For fixed u, v, we have

$$\begin{aligned} E(X_1) &= \Sigma\sigma^{mn}b_m a_n + u\Sigma\sigma^{mn}b_m b_n + v\Sigma\sigma^{mn}b_m c_n \\ &= A_1 + B_1 u + C_1 v \\ E(X_2) &= \Sigma\sigma^{mn}C_m a_n + u\Sigma\sigma^{mn}c_m b_n + v\Sigma\sigma^{mn}c_m c_n \\ &= A_2 + B_2 u + C_2 v \end{aligned}$$

and variances and covariance

$$\begin{aligned} \tau_{11} &= \Sigma\sigma^{mn}b_m b_n = B_1 \\ T_{12} &= \Sigma\sigma^{mn}b_m c_n = C_1 = B_2 \\ \tau_{22} &= \Sigma\sigma^{mn}c_m c_n = C_2. \end{aligned}$$

We may for example, estimate u and v by solving the equations

$$\begin{aligned} B_1 u + c_1 v &= X_1 - A_1 \\ B_2 u + C_2 v &= X_2 - A_2, \end{aligned}$$

or we may set up regions in the X_1, X_2 plane for which certain decisions are made. For example, when classifying an individual into one of three populations, we might delineate regions, as in Fig. 2.

Then the particular individual would be classified as coming from population I, II, or III, according to which region X_1, X_2 falls in. The individual points shown in the figure represent the expected values of X_1, X_2 for each of the three populations. No exhaustive investigation has been made for this situation, but some fairly obvious methods are available for constructing such regions.

With respect to significance tests when the σ_{mn}, a_m, b_m, c_m are estimated from samples, the whole gamut of multivariate analysis has to be run. Tests analogous to (but more complicated than) F tests exist for testing the significance

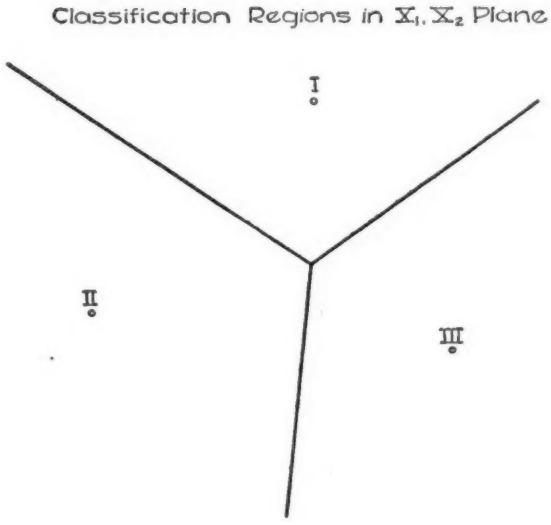


FIG. 2

of the discrimination, the significance of a subset of the x's, and the significance of a theoretical pair $X_{1,0}$, $X_{2,0}$ (Wilks [41], [42], [43]).

For some purposes a two-dimensional discrimant function X_1, X_2 may be unsatisfactory. For example, we might suspect that $\beta_1\gamma_2 = \beta_2\gamma_1$ (or that the relationship is nearly satisfied). Under these circumstances X_1 is (nearly) proportional to X_2, and we would like to compute the best one-dimensional discriminant function, even though we have started with two linear parameters u and v. Even if $\beta_1\gamma_2 \neq \beta_2\gamma_1$ we might still ask for the best one-dimensional discriminant function, in order to rank our populations on the "best" linear scale. If we define Y as that linear combination of $x_1, x_2, \cdots, x_p$ which has the largest multiple correlation with u and v, we have generalized the simple one-dimensional discriminant function in a second direction.

Before proceeding, it is useful to recognize that Y, as defined above, must be a

function of X_1 , X_2 , since X_1 and X_2 together contain all the information about u and v that can be obtained from the x's.

Now suppose we consider an arbitrary linear combination $Y = \lambda_1 X_1 + \lambda_2 X_2$. Y correlates best with

$$\lambda_1(\tau_{11}u + \tau_{12}v) + \lambda_2(\tau_{12}u + \tau_{22}v) = (\lambda_1\tau_{11} + \lambda_2\tau_{12})u + (\lambda_2\tau_{12} + \lambda_2\tau_{22})v.$$

We now have to choose λ_1 and λ_2 to maximize this correlation. This correlation will be maximized if we maximize the ratio of the variance of

$$(\lambda_1\tau_{11} + \lambda_2\tau_{12})u + (\lambda_1\tau_{12} + \lambda_2\tau_{22})v$$

(over the distribution of u and v values) to the variance of Y for fixed u and v. Call the first quantity S_1 , the second S_2 . Then $S_2 = \lambda_1^2\tau_{11} + 2\lambda_1\lambda_2\tau_{12} + \lambda_2^2\tau_{22}$ and S_1 is of the form $\lambda_1^2\mu_{11} + 2\lambda_1\lambda_2\mu_{12} + \lambda_2^2\mu_{22}$ where

$$\mu_{11} = \tau_{11}^2\sigma_{uu} + 2\tau_{11}\tau_{12}\sigma_{uv} + \tau_{12}^2\sigma_{vv}$$

$$\mu_{12} = \tau_{11}\tau_{12}\sigma_{uu} + (\tau_{12}^2 + \tau_{11}\tau_{22})\sigma_{uv} + \tau_{12}\tau_{22}\sigma_{vv}$$

$$\mu_{22} = \tau_{12}^2\sigma_{uu} + 2\tau_{12}\tau_{22}\sigma_{uv} + \tau_{22}^2\sigma_{vv} .$$

Maximizing S_1/S_2 leads to the equations:

$$\lambda_1\tau_{11} + \lambda_2\tau_{12} = \frac{S_1}{S_2}(\lambda_1\mu_{11} + \lambda_2\mu_{12})$$

$$\lambda_1\tau_{12} + \lambda_2\tau_{22} = \frac{S_1}{S_2}(\lambda_1\mu_{12} + \lambda_2\mu_{22})$$

i.e.

$$\lambda_1(\tau_{11} - \theta\mu_{12}) + \lambda_2(\tau_{12} - \theta\mu_{12}) = 0$$

$$\lambda_1(\tau_{12} - \theta\mu_{12}) + \lambda_2(\tau_{22} - \theta\mu_{22}) = 0, \qquad \text{with } \theta = S_1/S_2 .$$

It is thus seen that θ must satisfy the quadratic equation

$$(\tau_{11} - \theta\mu_{11})(\tau_{22} - \theta\mu_{22}) - (\tau_{12} - \theta\mu_{12})^2 = 0,$$

in order for solutions λ_1, λ_2 to exist. In general there will be two solutions, of which the greater corresponds to that linear combination $\lambda_1 X_1 + \lambda_2 X_2$ which has greatest multiple correlation with u and v, whereas the smaller corresponds to that linear combination which has least multiple correlation with u and v. θ itself corresponds to $R^2/(1 - R^2)$ for the regression of $\lambda_1 X_1 + \lambda_2 X_2$ on u, v.

In the general case with s degrees of freedom corresponding to w_1 , w_2 , $\cdots$, w_s , there is an s-dimensional discriminant function $(X_1$, X_2 , $\cdots$, $X_s)$, and a set of s linear combinations for which $R^2/(1 - R^2)$ is stationary with respect to

$$\lambda_1 , \cdots , \lambda_s .$$

The s roots (corresponding to an equation of degree s) arranged in decreasing order, permit construction of the best one-dimensional, two-dimensional, $\cdots$, $(s - 1)$-dimensional discriminant functions.

Discussion of the relevant significance tests for these reduced discriminant functions is beyond the scope of this paper. Reference may be made to the work of Hotelling and Fisher.

REFERENCES

[1] M. M. Barnard, "The secular variations of skull characters in four series of Egyptian skulls", *Ann. Eugen.*, Vol. 6 (1935), pp. 352–371.

[2] M. S. Bartlett, "The standard errors of discriminant function coefficients", *Jour. Roy. Stat. Soc.*, Suppl. 6 (1939), pp. 169–173.

[3] M. S. Bartlett, "Statistical significance of cannonical correlations", *Biometrika*, Vol. 32 (1942), pp. 29–37.

[4] W. D. Baten, "The discriminant function applied to spore measurements", *Mich. Acad. of Sci., Arts, and Letters*, Vol. 29 (1943), pp. 3–7.

[5] W. D. Baten and C. C. DeWitt, "Use of the discriminant function in the comparison of proximate coal analyses", *Indust. and Eng. Chem., Anal. Ed.*, Vol. 16 (1944), pp. 32–34.

[6] W. D. Baten and H. M. Hatcher, "Distinguishing method differences by use of discriminant functions", *Jour. of Exp. Ed.*, March, 1944.

[7] W. D. Baten, "The use of discriminant functions in comparing judges' scores concerning potatoes", *Jour. Amer. Stat. Assoc.*, Vol. 40 (1945), pp. 223–227.

[8] R. C. Bose, "On the exact distribution of D^2 statistic", *Sankhya*, Vol. 2 (1936), pp 143–154.

[9] R. C. Bose and S. N. Roy, "The exact distribution of the Studentized D^2statistic", *Sankhya*, Vol. 4 (1938) pp. 19–38.

[10] G. W. Brier, R. G. Schoot, and V. L. Simmons, "The discriminant function applied to quality rating in sheep", *Proc. Amer. Soc. An. Prod.*, Vol. 1 (1940), pp. 153–160.

[11] W. G. Cochran, "The comparison of different scales of measurement for experimental results", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 205–216.

[12] G. M. Cox and W. P. Martin, "Use of a discriminant function for differentiating soils with different Azotobacter populations", *Iowa State Col. Jour. of Sci.*, Vol. 11 (1937), pp. 323–331.

[13] B. B. Day and M. M. Sandomire, "Use of the discriminant function for more than two groups", *Jour. Amer. Stat. Assoc.*, Vol. 37 (1942), pp. 461–472.

[14] W. E. Deming, "On the chi test and curve fitting", *Jour. Amer. Stat. Assoc.*, Vol. 29 (1934), pp. 372–382.

[15] D. Durand, "Risk elements in consumer installment financing", *Nat. Bur. Econ. Res., Inc.*, Studies in Consumer Installment Financing, No. 8, 1941.

[16] R. A. Fisher, "The general sampling distribution of the multiple correlation coefficient", *Proc. Roy. Soc. A.*, Vol. 121 (1928), pp. 654–673.

[17] R. A. Fisher, "*Statistical Methods for Research Workers*", Oliver and Boyd, Section 29.

[18] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Ann. Eugen.*, Vol. 7 (1936), pp. 179–188.

[19] R. A. Fisher, "Statistical utilization of multiple measurements", *Ann. Eugen.*, Vol. 8 (1938), pp. 376–386.

[20] R. A. Fisher, "The sampling distribution of some statistics obtained from non-linear equations", *Ann. Eugen.*, Vol. 9 (1939), pp. 238–249.

[21] R. A. Fisher, "The precision of discriminant functions", *Ann. Eugen.*, Vol. 10 (1940), pp. 422–429.

[22] H. E. Garrett, "The discriminant function and its use in psychology", *Psychometrika*, Vol. 8 (1943), pp. 65–79.

[23] M. A. Girshick, "On the sampling theory of the roots of determinantal equations", *Annals of Math. Stat.*, Vol. 10 (1939), pp. 203–224.

[24] T. Haalvelmo, "Probability applications in econometrics", *Econometrica*, Suppl. Vol. 12 (1944).

[25] H. Hotelling, "Generalization of Student's ratio", *Annals of Math. Stat.* ,Vol. 2 (1931), pp. 360–378.

[26] H. Hotelling, "The most predictable onterion", *Jour. Educ. Psychol.*, Vol. 26 (1935), pp. 139–142.

[27] H. Hotelling, "Relations between two sets of variates", *Biometrika*, Vol. 28 (1936), pp. 321–377.

[28] P. L. Hsu, "On the distribution of roots of certain determinantal equations", *Ann. Eugen.*, Vol. 9 (1939), pp. 250–258.

[29] W. G. Madow, "Contributions to the theory of multivariate statistical analyses", *Trans. Am. Math. Soc.*, Vol. 44 (1938), pp. 454–495.

[30] P. C. Mahalanobis, "Analysis of race mixture in Bengal", *Jour. Asiat. Soc. Beng.*, Vol. 23 (1927), pp. 301–333.

[31] P. C. Mahalanobis, "On tests and measures of group divergence", *Jour. Asiat. Soc. Beng.*, Vol. 26 (1930), pp. 541–588.

[32] P. C. Mahalanobis, "On the generalized distance in statistics", *Proc. Mat. Inst. Sci. Ind.*, Vol. 12 (1936), pp. 49–55.

[33] E. A. Martin, "A study of an Egyptian series of mandibles, with special reference to mathematical methods of sexing", *Biometrika*, Vol. 28 (1936), pp. 149–178.

[34] K. Pearson, "On the coefficient of racial likeness", *Biometrika*, Vol. 18 (1926) ,pp. 105–117.

[35] C. R. Rao, "Tests with discriminant functions in multivariate analysis", *Sankhya*, Vol. 7 (1946), pp. 407–414.

[36] H. F. Smith, "A discriminant function for plant selection", *Ann. Eugen.*, Vol. 7 (1936), pp. 240–250.

[37] W. L. Stevens, "The standardization of rubber flexing tests", *India Rubber World*, August 1, 1940.

[38] A. Wald, "On a statisical problem arising in the classification of an individual into one of two groups", *Annals of Math. Stat.*, Vol. 14 (1944), pp. 145–162.

[39] N. Wallace and R. M. W. Travers, "A psychometric sociologia study of a group of specialty salesmen", *Ann. Eugen.*, Vol. 8 (1938), pp. 266–302.

[40] F. V. Waugh, "Regression between sets of variables", *Econometrica*, Vol. 10 (1942), pp. 290–310.

[41] S. S. Wilks, "Certain generalizations in the analysis of variance", *Biometrika*, Vol. 24 (1932), pp. 471–494.

[42] S. S. Wilks, "On the independence of k sets of normally distributed statistical variables", *Econometrica*, Vol. 3 (1935), pp. 309–326.

[43] S. S. Wilks, *Mathematical Statistics*. Princeton Univ. Press, 1943.

NON-PARAMETRIC ESTIMATION II. STATISTICALLY EQUIVALENT BLOCKS AND TOLERANCE REGIONS—THE CONTINUOUS CASE

By John W. Tukey

Princeton University

1. Summary. Wald [2, 1943] extended the usefulness of tolerance limits to the simplest multi-dimensional cases. His principle is here used to provide many new ways of using a sample of n to divide the range of the population into $n + 1$ blocks of known behavior. The exact tolerance distribution for the proportions of the population covered by these blocks is extended from the case of a continuous probability density function to the case of a continuous cumulative distribution function. Such an extension is needed in dealing completely with multivariate cases *even* where the underlying distribution is as smooth as a multivariate normal distribution.

The devices used in Paper I [1] to extend the usefulness of tolerance limits to the case of a discontinuous underlying distribution will be applied in the next paper of this series, with some extension, to extend the usefulness of these general tolerance regions to the case of a discontinuous distribution. Some of these results specialize into new results for the univariate case, although they do not seem to have any immediate practical application.

The author wishes to acknowledge the stimulation given to his work on this problem by Henry Scheffé, whose modesty has kept this paper from the joint authorship of papers I [1, Scheffé and Tukey 1945] and IV (not yet written).

2. Introduction. Wald's great contribution to the theory of tolerance limits was his method of successive elimination. As originally presented for a bivariate situation it ran roughly as follows: Let (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) be a sample of n from an arbitrary bivariate population. The type of tolerance region to be used is determined by four preassigned integers, k_1, k_2, k_3, and k_4. The procedure is as follows: Order the n observations according to their x values. Select the k_1 highest, and let the x coordinate of the lowest of these k_1 be x_u. Select the k_2 lowest, and let the x coordinate of the highest of these k_2 be x_l. Discard these $k_1 + k_2$ selected observations, and order the remaining $n - k_1 - k_2$ observations according to their y values. Select the k_3 highest of these remaining observations, and let the y coordinate of the lowest of these k_3 be y_u. Select the k_4 lowest of these remaining observations, and let the y coordinate of the highest of these k_4 be y_l. The tolerance region, consisting of all points (x, y), with $x_l < x < x_u$ and $y_l < y < y_u$ depends on the sample, and, hence, so does the fraction of the population falling in (= covered by) this region. Wald showed that the distribution of this fraction covered was independent of the underlying bivariate distribution, so long as this latter distribution had a continuous probability density function. He showed that the

distribution was the same as that arising in the one-dimensional case when a tolerance region was set with the aid of $k_1 + k_2 + k_3 + k_4$ observations. (Numerical approximation to these distributions will be discussed in Paper IV of this series.

The important device in this process, and the one which makes the conclusion possible, is the discarding of the $k_1 + k_2$ observations after they have played their part by determining x_l and x_u.

We shall shortly be able to describe this procedure of Wald's as a special case of a more general procedure, but we shall first go back to the simplest one dimensional case to explain some of our notions and terminology.

Consider the uniform distribution from 0 to 1, draw a sample of n, and let the sample values, ordered according to size be $t_1, t_2, \cdots, t_n$. These n values divide the interval from 0 to 1 into the following $n + 1$ parts $(0, t_1), (t_1, t_2), \cdots, (t_{n-1}t_n), (t_n, 1)$ which we shall call *blocks*. Since the joint distribution of the t_i is well known, that of the lengths of these $n - 1$ blocks is easily found. This distribution of lengths would be unimportant, if it were not at the same time the distribution of the fractions of the population covered by the blocks. As is shown later, this distribution of fractions covered, or, more simply, of *coverages*, has the following properties:

(i) the fractions covered add up to 1.

(ii) the distribution is completely symmetrical.

Property (ii) makes intuitive the result of Wilks [3, 1941] that the distributions of the coverage of regions obtained

(a) by removing the $k_1 + k_2$ left-most blocks,

(b) by removing the k_1 left-most and the k_2 right-most blocks

are identical. The specific distribution obtained satisfies

(iii) if the coverages are taken as barycentric coordinates on an n-simplex, the distribution over the simplex is uniform,

(iv) the sum of the coverages of any k *pre*selected blocks of the $n + 1$ has the well-known distribution

$$Pr\{\text{sum of } k \text{ coverages} < t\} = I_t(n - k + 1, k)$$

where $I_\beta(n, m)$ is the incomplete Beta function.

We shall call a set of blocks, derived from a sample, whose coverages behave in this general way a set of *statistically equivalent blocks*. Normally this will be abbreviated to se-blocks. (A precise definition is given in section 4.)

We shall concentrate much of our attention on all the blocks and their symmetrical character, rather than on the tolerance region formed by deleting k of them, since our results will then be applicable to many other problems.

Now we can generalize Wald's original procedure. Let $W_1, W_2, \cdots, W_n$ be a sample of n—we shall not need to consider its distribution—and let $\varphi_1, \varphi_2, \cdots, \varphi_n$ be n numerically valued functions of W, possibly alike, possibly distinct, such that $\varphi_1(W), \varphi_2(W), \cdots, \varphi_n(W)$ have a joint distribution. Proceed as follows:

Order the W_i according to the numbers $\varphi_1(W_i)$, select the W_i for which $\varphi_1(W_i)$ is largest and denote it by $W_{i(1)}$. The first block contains all W such that

$$\varphi_1(W) > \varphi_1(W_{i(1)}). \tag{2.1a}$$

Discarding $W_{i(1)}$, order the remaining W_i according to the values of $\varphi_2(W_i)$, and select as $W_{i(2)}$ the one giving the largest value. The second block contains all W such that

$$\begin{aligned} \varphi_1(W) &< \varphi_1(W_{i(1)}), \\ \varphi_2(W) &> \varphi_2(W_{i(2)}). \end{aligned} \tag{2.1b}$$

Continue this process. The mth block, for $m \leq n$ will be defined by

$$\begin{aligned} \varphi_j(W) &< \varphi_j(W_{i(j)}), \qquad j = 1, 2, \cdots, m-1, \\ \varphi_m(W) &> \varphi_m(W_{i(m)}), \end{aligned} \tag{2.1m}$$

and the $(n+1)$st block by

$$\varphi_j(W) < \varphi_j(W_{i(j)}), \qquad j = 1, 2, \cdots, n. \tag{2.1n}$$

(A graphical example of this construction is given shortly.) This set of $n+1$ blocks will be statistically equivalent whenever the cumulative distribution of each φ_i function is continuous.

To specialize this to the case described above, let W be a pair (x, y) of numbers and let

(i) the first k_1 φ's be the x-coordinate of W,
(ii) the next k_2 φ's be *minus* the x-coordinate of W,
(iii) the next k_3 φ's be the y-coordinate of W,
(iv) the next k_4 φ's be *minus* the y-coordinate of W,
(v) the remaining φ's be arbitrary.

Then the first k_1 blocks will contain all W for which

$$x = \varphi_j(W) > \varphi_j(W_{i(j)}), \qquad j = 1, 2, \cdots, k_1$$

that is, for which

$$x > x_u = \varphi_{k_1}(W_{i(k_1)}).$$

Similarly, the next $k_2 + k_3 + k_4$ blocks will contain all W with

$$\begin{aligned} x &< x_l, \\ y &> y_u, \qquad x_l \leq x \leq x_u, \\ y &< y_l, \qquad x_l \leq x \leq x_u, \end{aligned}$$

respectively, and the removal of these $k_1 + k_2 + k_3 + k_4$ blocks leaves Wald's tolerance region (plus the boundaries where $x = x_u$, $x = x_l$, $y = y_u$, $y = y_l$).

There would be no point in this more general wording, if it did not include

new cases of some interest. We give now, in graphic terms, an example of such a case.

We deal with a sample of n bivariate observations, which we think of as plotted on a *map* so that we can use geographical language. The number n is rather large, and we wish to construct a tolerance region by deleting 12 blocks. We proceed as follows:

Find the most northerly point, draw an East-West line through it, and shade the area North of the line. Find the most easterly point in the unshaded area, draw a North-South line through it, and shade the unshaded area East of the

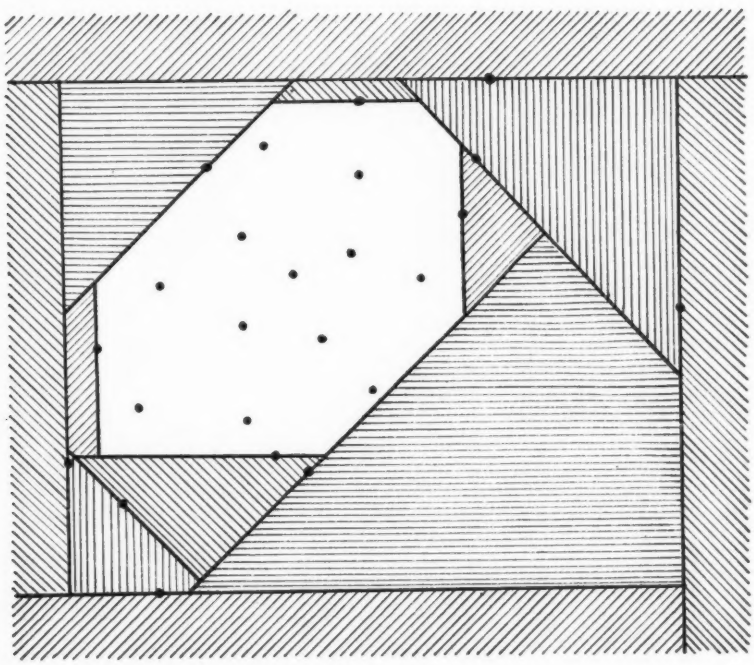


FIG. 1

line. Find the most southerly point, (always working in the unshaded area), draw an East-West line through it and shade the area South of the line. Find the most westerly point, draw a North-South line through it, and shade the area West of the line. Find the most northeasterly point, draw a NW-SE line through it and shade the area northeast of the line. Find the most southeasterly point, draw a NE-SW line through it, and shade the area southeast of the line. Repeat this 6 times more, choosing in succession the most southwesterly, northwesterly, northerly, easterly, southerly, and westerly points. The remaining points will now lie in an unshaded area surrounded by a polygon, which will have 8 (or perhaps fewer) sides. The inside of this polygon is the desired tolerance region.

Figure 1 shows the final result, starting from $n = 25$. The practicing statistician is invited to try an example of his own with n at least 100.

Other newly accessible cases can easily be invented by the reader, after he considers this example carefully.

The use of a single W and n functions φ_i has two virtues; it simplifies notation and frees the intuition, as compared with the use of n chance quantities $Z_i = \varphi_i(W)$.

If the bivariate situation above were regarded as a 12-variate situation, where the variates were, in order, $(y, x, -y, -x, x + y, x - y, -x - y, -x + y, y, x, -y, -x)$ then the original Wald procedure with $k_1 = k_3 = \cdots = k_{23} = 1$; $k_2 = k_4 = \cdots = k_{24} = 0$ would apply to construct the same region. Yet even if x and y had a bivariate normal distribution, Wald's proof would not apply without extension. For the 12-dimensional distribution is highly singular (it is concentrated on a 2-dimensional plane in 12-dimensional space) and there is no hope of a density function. An extension of Wald's result to the case where the 12-dimensional joint cumulative distribution function is continuous —as is the case in this example when x and y have a continuous joint cumulative —is clearly needed.

When we come to deal with the case of where the cumulative needs not be continuous we shall meet a further difficulty, namely "ties". But if, as in the present case, the cumulative is continuous, it is easy to see that the probability that $\varphi_i(W_j) = \varphi_i(W_k)$ for any i, j, k is zero.

3. Terminology and notation. A quantity which has a probability distribution we call a *chance quantity* (it has frequently been called a *random variable*). The term chance quantity does not imply that its values are single real numbers, they may be single real numbers (when we also speak of a real chance quantity), sets of n real numbers, or more general objects. The cumulative distribution function, or *cumulative*, of a single real chance quantity, X, is defined by

$$F(t) = Pr\{X < t\},$$

except perhaps at the discontinuities of F. We have used here the notation $Pr\{k(X)\}$ to indicate the probability that $k(X)$ holds, and we have followed our policy of using capital letters for chance quantities and the corresponding lower case letters for their values.

The set of values of W, or, as we shall say, the W-set, for which, for example $\varphi(W) \leq 3$, will be denoted by

$$\{W \mid \varphi(W) \leq 3\}.$$

We shall wish to compute probabilities associated with one or more functions of a chance quantity; usually we will emphasize that these functions shall be measurable with respect to the probability measure underlying the distribution of W by asserting that they have a joint cumulative, which is defined by

$$F(t_1, t_2, \cdots, t_k) = Pr\{\varphi_k(W) < t_k\},$$

(except possibly at discontinuities of F) and which does not exist unless the φ_i are measurable with respect to the unknown underlying distribution of W. In cases where we neglect to remind the reader, it is still assumed that the functions are measurable.

The coverage of a W-set, which may itself be a chance quantity, is defined by

$$\text{Coverage of } S = Pr\{W \in S\}.$$

When S is a chance quantity, its coverage is also a chance quantity. The barycentric simplex (of dimension n) is the set of points in $n+1$-dimensional Euclidean space $(t_1, t_2, \cdots, t_{n+1})$ with $t_1 + t_2 + \cdots + t_{n+1} = 1$ and $0 \leq t_i \leq 1$. The name comes from the representation of the point $(t_1, t_2, \cdots, t_{n+1})$ as the center of gravity (in mechanical terms) or mean (in statistical terms) of the distribution where a fraction t_i is concentrated at the ith vertex. (In order, the vertices are $(1, 0, 0, \cdots, 0)$, $(0, 1, 0, \cdots, 0)$, etc.) The uniform distribution on this simplex has an (n-dimensional) density

$$n!dt_1dt_2 \cdots dt_n, \qquad (0 \leq t_1, t_2, \cdots, t_n, 1 - t_1 - t_2 \cdots - t_n \leq 1),$$

and the cumulative

$$T(x_1, x_2, \cdots, x_{n+1}) = n! \int\int \cdots \int dt_1 dt_2 \cdots dt_n$$

where the integration is over the range where $0 \leq t_i \leq x_i$ *and at the same time* $t_1 + t_2 + \cdots + t_{n-1} \leq 1$.

4. The blocks determined by n values of W. We deal now with a population of W's (a probability measure μ on the space $T \equiv \{w\}$), a family of functions $\varphi_1, \varphi_2, \cdots, \varphi_m$ of W with a joint cumulative (measurable with respect to μ) and a set of values $w_1, w_2, \cdots, w_n$, $(w_i \in T)$.

(4.1) DEFINITION *The set* $w_1, w_2, \cdots, w_n$ *and the functions* $\varphi_1, \varphi_2, \cdots, \varphi_m$ *define blocks as follows:*

$$S_1 = \{w \mid \varphi_1(w) > a_1\} \tag{4.2}$$

where $a_1 = \max_i \varphi_1(w_i) = \varphi_1(w_{i(1)})$, *which defines* $i(1)$.

$$S_2 = \{w \mid \varphi_1(w) < a_1, \varphi_2(w) > a_2\}, \tag{4.3}$$

where $a_2 = \max_{i \neq i(1)} \varphi_2(w_i) = \varphi_2(w_{i(2)})$, $i(2) \neq i(1)$, *which defines* $i(2)$. *And in general, for* $1 < k \leq \min(m, n)$,

$$S_k = \{w \mid \varphi_1(w) < a_1, \cdots, \varphi_{k-1}(w) < a_{k-1}, \varphi_k(w) > a_k\}, \tag{4.4}$$

where $a_k = \max_i{}' \varphi_k(w_i) = \varphi_k(w_{i(k)})$, *the maximum being taken over all* i except $i(1), i(2), \cdots, i(k-1)$; and $i(k)$ *being chosen distinct from all* $i(j), j < k$.

If $m \geq n$, *then*

$$S_{n+1} = \{w \mid \varphi_1(w) < a_1, \cdots, \varphi_n(w) < a_n\}. \tag{4.5}$$

If $m \leq n$, *then*

$$S_{m|n+1} = \{w \mid \varphi_1(w) < a_1, \cdots, \varphi_m(w) < a_m\}. \tag{4.6}$$

The result of this definition is to use $w_1, \cdots, w_n$ and $\varphi_1, \cdots, \varphi_m$ to define $n+1$ blocks (one more than there are w's) in case there are enough functions, and, in case there are not enough functions, to define one small block, S_i, for each function plus one large remainder $S_{m|n+1}$. We notice

(4.2) REMARK. *The blocks of* (4.1) *are well defined unless* $\varphi_1(w_j) = \varphi_i(w_k)$ *for some* i, j, k.

5. Statement of results for the statistician. The central results can be stated as follows:

(5.1) THEOREM $A_{m|n+1}$. *If* $W_1, W_2, \cdots, W_n$ *are a sample of* n *from a distribution, if* $\varphi_1, \varphi_2, \cdots, \varphi_m$, $(m \leq n)$, *are* m *functions such that*

$$\varphi_1(W), \varphi_2(W), \cdots, \varphi_m(W)$$

have a joint distribution which has a continuous cumulative, and if the blocks $S_1, S_2, \cdots, S_m$ *and* $S_{m|n+1}$ *are defined as in* (4.1), *then*

(i) *the blocks are disjoint chance sets, uniquely defined with probability one,*

(ii) *the distribution of the coverages*

$$c_i = Pr\{w \text{ in } S_i\}, \qquad i = 1, 2, \cdots, m$$

and

$$c_{m|n+1} = Pr\{w \text{ in } S_{m|n+1}\}$$

is the same as that of $t_1, t_2, \cdots, t_m$ *and* $t_{m+1} + t_{m+2} + \cdots + t_{m+1}$ *where* t_i *are uniformly distributed on the barycentric simplex with* $n+1$ *vertices.*

Conditions (5.1i) and (5.1ii) are the precise definition of *a partial family of statistically equivalent blocks of type* $n+1$ *and an associated* $(m \mid n+1)$ *tolerance region.*

(5.2) THEOREM B_{n+1}. *If* $W_1, W_2, \cdots, W_n$ *are a sample of* n *from a distribution, and if* $\varphi_1, \varphi_2, \cdots, \varphi_m$, $(m \geq n)$, *are* m *functions such that*

$$\varphi_1(W), \varphi_2(W), \cdots, \varphi_m(W)$$

have a joint distribution which has a continuous cumulative, and if the blocks $S_1, S_2, \cdots, S_{n+1}$ *are defined as in* (4.1), *then*

(i) *the blocks are disjoint chance sets, defined with probability one.*

(ii) *the distribution of the coverages*

$$c_i = Pr\ \{w \text{ in } S_i\}, \qquad i = 1, 2, \cdots, n+1$$

is the same as that of $t_1, t_2, \cdots, t_{n+1}$, *where the* t_i *are uniformly distributed on the barycentric simplex with* $n+1$ *vertices.*

Conditions (5.2i) and (5.2ii) are the precise definition of *a complete family of statistically equivalent blocks.* In Paper III we shall have to widen these notions a little, and this form will then be qualified by the phrase "in the narrow sense".

6. Statement of results for the measure theorist. The construction of (4.1) maps the product $T^n \times U^n$ into E^{n+1} where T is the set of w's (and hence T^n is the set of ordered n-tuples of w's), U is the space of all real-valued functions defined over T, measurable with respect to a fixed probability measure μ, and possessing a continuous cumulative, (i.e. $\mu(\{w \mid \varphi(w) = c\}) = 0$ for all real c), and hence U^n is the space of ordered n-tuples of such functions, and E^{n+1} is Euclidean n-dimensional space. More precisely, the mapping is into the barycentric simplex with $n+1$ vertices, a subset of E^{n+1}, and is well defined except for a set in T^n of measure zero with respect to μ^n, the power measure of μ. In these terms, we may restate theorem B as follows:

(6.1) THEOREM B_{n+1}. *Hold the n functions* $\varphi_1, \varphi_2, \cdots, \varphi_n$ *and the probability measure fixed, then* T^n *is mapped into* B_n *and the power measure* μ^n *is carried by that mapping into a measure on* B_n. *This measure is always* $n!$ *times Lebesgue measure.*

7. Wald's principle. The essential principle behind Wald's process of discarding observations is sufficiently fundamental to warrant a name of its own. It can be stated, quite generally, in the two following forms:

(7.1) WALD'S PRINCIPLE. (discrete form.) *Let W be a chance quantity, and consider samples of n. Fix disjoint w-sets* $A_1, A_2, \cdots, A_m, B$. *Consider those samples of n for which exactly one value falls in each* A_i *and the remaining n-m fall in B. The distribution of the n-m falling in B is that of a random sample of n-m from the distribution of W restricted to B.* (i.e. $\mu_B(X) = [\mu(B)]^{-1}\mu(BX)$.)

(7.2) WALD'S PRINCIPLE. (conditional form.) *Let W be a chance quantity, and* φ *a function such that each value of* $\varphi(W)$ *has probability zero. Consider samples of n. Then the conditional distribution of the* w_i, *given that*

$$\max_i \varphi(w_i) = a,$$

is that of one w_{io} *with* $\varphi(w_{io}) = a$ *and a sample of* $n-1$ *other* w_i *from the distribution of W restricted to* $B = \{w \mid \varphi(w) < a\}$.

(7.3) CENTRAL LEMMA. *Let W be a chance quantity and let* $\varphi_1, \cdots, \varphi_n$ *be functions with a joint cumulative such that* $\varphi_i(w) = a$ *has probability zero for each i and a (i.e. the joint cumulative is continuous). Then the conditional distribu-*

tion of the remaining $n-k$ w's, after k blocks have been chosen according to (4.1) *is that of a sample from the distribution of W restricted to*

$$B = \{w \mid \varphi_1(w) < a_1, \cdots, \varphi_k(w) < a_k\},$$

where $k = 1, 2, \cdots, n$.

The proofs of these statements are elementary and direct. To establish (7.1) we have only to show that given two sets in B^{n-k}, their probabilities on the assumption that one w_i is in each A_i are in the ratio of their probabilities for an unrestricted sample of $n-k$. But the probability of finding the $n-k$ w_i in a set R, contained in B^{n-k}, and one w_i in each A_i, is exactly

$$\frac{n!}{(n-k)!}\mu(A_1)\mu(A_2)\cdots\mu(A_k)$$

times the probability that $n-k$ w_i, known to be in B^{n-k}, will fall in R. This establishes (7.1).

In order to prove (7.2) we must show that the probability of a set R of n-tuples $w_1, w_2, \cdots, w_n$ is the same whether calculated directly or calculated by the proposed conditional distribution. To this end, it is natural to decompose R as follows:

$$R = R(1) + R(2) + \cdots + R(n) + Z,$$

where $R(i)$ contains those $(w_1, \cdots, w_n)$ in R for which $\varphi(w_i) > \varphi(w_j)$ for all $j \neq i$, and Z contains the remaining $(w_1, \cdots, w_n)$; which must involve at least one tie $\varphi(w_j) = \varphi(w_k)$, $j \neq k$. Since Z has probability zero, it will suffice to establish the equality of the two calculations for sets of the form $R(i)$, and because of symmetry we may restrict ourselves to sets of the form $R(1)$.

Given an integer N, we decompose the range of $\varphi(w)$ into Nn segments of equal probability, which we may do because the cumulative of φ is continuous. There are then Nn values b_k, $(b_o = -\infty, b_{Nn} = +\infty)$ such that

$$Pr\{b_{k-1} < \varphi(w) < b_k\} = 1/Nn.$$

We now decompose our set R (which is of the form $R(1)$ as follows:

$$R = R_2 + \cdots + R_{Nn} + Y,$$

where R_k contains those n-tuples

$$(w_1, \cdots, w_n) \text{ for which } b_{k-1} < \varphi(w_1) < b_k$$

and $\varphi(w_i) < b_{k-1}$ for all $i > 1$. The remaining set Y contains n-tuples where the two largest $\varphi(w_i)$, $(i = 1 \text{ and } i = i_o)$, belong to the same interval. The probability of this is less than

$$\frac{n(n-1)}{2}\left(\frac{1}{nN}\right)^2 \leq \frac{1}{2N^2}$$

as calculated from the known distribution. Calculating from the conditional distribution, we find immediately a bound of

$$\sum_k \left\{\left(\frac{k}{Nn}\right)^n - \left(\frac{k-1}{Nn}\right)^n\right\} \frac{n-1}{k} = \frac{n-1}{(Nn)^n} \sum \frac{k^n - (k-1)^n}{k}$$

$$= \frac{n-1}{(Nn)^n}\left((nN)^{n-1} + \sum_{k<nN} \frac{k^{n-1}}{k-1}\right)$$

$$\leq \frac{n-1}{(Nn)^n} (A_n (nN)^{n-1})$$

$$= \frac{A_n}{N},$$

where A_n is a constant depending only on n. Thus, as N increases, the probability of the successive sets Y tend to zero—calculated either way. To show the equivalence of the two calculations it is now sufficient to show that they agree for the sets R_k. But this is a case of (7.1) and the lemma is proved.

Now (7.3) follows by induction, applying (7.2) at each step.

8. Proof of theorems. We notice that Theorem B_n is equivalent to Theorem $A_{m|n+1}$, since, according to (4.1) $S_{n|n+1} = S_{n+1}$.

We have only to prove theorem $A_{m|n+1}$, which we do by induction on m. For $m = 1$, it is exactly Wilks' [3, 1941] original one-dimensional theorem, and is known. Let us assume it for $m = k$ and demonstrate it for $m = k + 1$, for by induction this will complete the proof.

We must deal with the blocks $S_1, S_2, \cdots, S_k, S_{k+1}$ and $S_{k+1|n+1}$, (notation as in (4.1) and (5.1)). We need the obvious

(8.1) LEMMA. *Since the cumulative of φ_{k+1} is continuous, the union of S_{k+1} and $S_{k+1|m+1}$ differs from $S_{k|m+1}$ by a set of zero probability.*

Hence

$$c_{k|n+1} \equiv c_{k+1} + c_{k+1|n+1}.$$

Since we know from the induction hypothesis that $c_1, c_2, \cdots, c_k$ and $c_{k|n+1}$ have the correct joint distribution, we have only to show that c_{k+1} and $c_1, c_2, \cdots, c_k$ have the correct joint distribution. Fix $c_1, c_2, \cdots, c_k$. Then $a_1, a_2, \cdots, a_k$ must be fixed, and so (7.3) applies to the $n-k$ w_i's not discarded after $a_1, a_2, \cdots, a_k$ have been fixed. The conditional distribution of c_{k+1} must be that of a fixed number $(1 - c_1 - c_2 - \cdots - c_k)$, which is the probability attached to $S_{k|n+1}$, times the coverage of one block based on a sample of $n-k$, since the remaining $n-k$ w's behave like a sample.

Consider the very particular case where w is uniformly distributed between zero and one and $\varphi_i(w) \equiv w$, all that we have said in the last paragraph applies —the conditional distribution of c_{k+1} given $c_1, c_2, \cdots, c_k$ is the same in the two cases—hence the joint distribution of $c_1, c_2, \cdots, c_k, c_{k+1}$ is the same in both cases—but in this very particular case the joint distribution is known to be that required by theorem $A_{k+1|n+1}$.

REFERENCES

[1] H. SCHEFFÉ AND J. W. TUKEY, "Non-parametric estimation. I. Validation of order statistics", *Annals of Math. Stat.*, Vol. 16 (1945), pp. 187–192. (Also cited as Paper I.)

[2] A. WALD, "An extension of Wilks' method for setting tolerance limits", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 45–55.

[3] S. S. WILKS, "Determination of sample sizes for setting tolerance limits", *Annals of Math. Stat.*, Vol. 12 (1941), pp. 91–96.

SOME BASIC THEOREMS FOR DEVELOPING TESTS OF FIT FOR THE CASE OF THE NON-PARAMETRIC PROBABILITY DISTRIBUTION FUNCTION, I

By Bradford F. Kimball

State Department of Public Service, New York, N. Y.

1. Summary. In developing tests of fit based upon a sample $O_n(x_i)$ in the case that the cumulative distribution function $F(X)$ of the universe of X's is not necessarily a function of a finite number of specific parameters—sometimes known as the non-parametric case—it has been pointed out by several writers that the "probability integral transformation" is a useful device (cf. [1]–[4]).

The author finds that a modification of this approach is more effective. This modification is to use a transformation of ordered sample values x_i from a random sample $O_n(x_i)$ based on successive *differences* of the cdf values $F(x_i)$.

A theorem is proved giving a simple formula for the expected values of the products of powers of these differences, where all differences from 1 to $n + 1$ are involved in a symmetrical manner.

The moment generating function of the test function defined as the sum of m squares of these successive differences is developed and the application of such a test function is briefly discussed.

2. Introduction. Let the sample values x_i be ordered so that

$$x_i \leqq x_{i+1}, \qquad (i = 1, 2, \cdots, n-1). \tag{2.1}$$

Let F_r denote the value of the cdf $F(X)$ associated with the rth ordered sample value x_r. Thus

$$F_r = F(x_r). \tag{2.2}$$

Consider the following transformation of the ordered sample values x_i based upon the (hypothetically) known cumulative distribution function $F(X)$ which will be taken as a continuous function of X over its admissible range:

$$\begin{aligned} u_1 &= F_1, \\ u_r &= F_r - F_{r-1}, \qquad (r = 2, 3, \cdots, n) \\ u_{n+1} &= 1 - F_n. \end{aligned} \tag{2.3}$$

The restrictions on F_i are that

$$F_i \leqq F_{i+1}, \text{ and } 0 \leqq F_i \leqq 1. \tag{2.4}$$

The above transformation (2.3) translates these conditions into the *symmetrical* conditions

$$0 \leqq u_i, \text{ and } u_1 + u_2 + \cdots + u_n + u_{n+1} = 1. \tag{2.5}$$

A one-to-one correspondence between u_i and F_i exists if one of the u_i be omitted,—say u_β. With u_β omitted, the Jacobian of the transformation from F_i to u_i

has value unity. The probability density of the sample $O_n(x_i)$, with x_i ordered, is given by

$$P[O_n(x_i)]\, dO_n = n!\, dF_1\, dF_2 \cdots dF_n . \tag{2.6}$$

Hence with u_β omitted,

$$P[O_n(x_i)]\, dO_n = n!\, du_1\, du_2 \cdots du_{\beta-1}\, du_{\beta+1} \cdots du_{n+1} . \tag{2.7}$$

The sample space of the u_i with u_β omitted, is that portion of the $n + 1$ Euclidean space of all the u_i variables, bounded by the coordinate hyperplanes, which is on the projection of the hyperplane (2.5) upon the hyperplane $u_\beta = 0$. This is a region in the n-space of the u_i with u_β omitted, bounded by the coordinate hyperplanes and the hyperplane

$$u_1 + u_2 + \cdots + u_{\beta-1} + u_{\beta+1} + \cdots + u_n + u_{n+1} = 1. \tag{2.8}$$

Thus the formal integral of the pdf of the u_i over sample space is

$$n! \int n \int du_1 \cdots du_{\beta-1}\, du_{\beta+1} \cdots du_{n+1} = 1 \tag{2.9}$$

with $0 \leqq u_i$, and u_i bounded above by the hyperplane (2.8).

It is now clear that both the pdf and the sample space of the u_i (with u_β omitted) are symmetrical in the u_i. This fact leads to *complete* symmetry of the joint distribution function of any set of u_i, over $i = 1$ to $n + 1$ including u_β, relative to the u_i selected. Other interesting results are forthcoming.

3. Basic mathematical theorem. Using the techniques associated with the Beta function, the expectation of the products of powers u_i is found to be

$$\begin{aligned} &E[u_r^p \cdot u_s^q \cdot u_t^w \cdots] \\ &= \Gamma(n+1)\Gamma(p+1)\Gamma(q+1)\Gamma(w+1) \cdots / \Gamma(n+p+q+w+\cdots+1) \end{aligned} \tag{3.1}$$

where r, s, t, etc., are any set of different indices (for the present other than β) from the integers 1 to $n + 1$, and p, q, w, etc., are any real numbers greater than minus one. The relation (3.1) can further be generalized to the case where u_β may be included. This will be proved for the case $n = 2$, with p, q and w taken as integers. The generalization can be concluded from inspection. Thus with

$$u_3 = 1 - u_1 - u_2,$$

$$\begin{aligned} E[u_1^p \cdot u_2^q \cdot u_3^w] &= 2! \int_0^1 u_2^q\, du_2 \int_0^{1-u_2} u_1^p (1 - u_1 - u_2)^w\, du_1 \\ &= 2! \int_0^1 u_2^q (1 - u_2)^{p+w+1} \int_0^1 s^p (1 - s)^w\, ds \\ &= \frac{2!\,p!\,w!}{(p+w+1)!} \int_0^1 u_2^q (1 - u_2)^{p+w+1}\, du_2 = \frac{2!\,p!\,q!\,w!}{(p+q+w+2)!} . \end{aligned}$$

Hence the theorem:

THEOREM. *Given a random sample of n values of X from a universe with* cdf *$F(X)$ which is continuous over the range of X. With the sample values x_i ordered so that $x_i \leqq x_{i+1}$ define a set of $n + 1$ variables u_i as the successive differences of $F(x_i)$ by the relations* (2.3). *The expected value of the product of real powers greater than minus one of any or all of the u_i, $(i = 1, 2, \cdots, n + 1)$, is given by the relation* (3.1) *above (not subject to the omission of u_β).*

There are many interesting consequences of this theorem. Perhaps the most striking is the following:

COROLLARY 1. *Let a range $\alpha(m, k)$ for positive integer m be defined by*

$$\alpha(m, k) = F(x_{k+m}) - F(x_k) \tag{3.2}$$

with $$k = 0, 1, 2, \cdots, n, \text{ and } m \leqq n + 1 - k$$

under the convention

$$F(x_0) = 0, \qquad F(x_{n+1}) = 1.$$

The probability distribution of $\alpha(m, k)$ is independent of k and hence is the same as that of $F(x_m)$.

Another interesting consequence (not new) is the following:

COROLLARY 2. *The correlation of u_i and u_k, $i \neq k$, is the same for all pairs (i, k) over the range of indices from 1 to $n + 1$, and has the value $-1/n$.*

Introducing the notation

$$[n + r]_r = (n + r)(n + r - 1) \cdots (n + 1), \tag{3.3}$$

the corollary follows from the relationships

$$E(u_i) = 1/(n + 1), \quad E(u_i^2) = 2/[n + 2]_2, \quad E(u_i u_k) = 1/[n + 2]_2.$$

The fact that the correlation between any two frequency differences u_i and u_k is negative leads to the following more general relationship:

COROLLARY 3. *For any set of different indices i, j, k, etc., and for any positive numbers p, q, r, etc., the expectation of the product of the powers $p, q, r, \cdots$ of $u_i, u_j, u_k \cdots$ is less than the product of the expectations of the powers taken separately:*

$$E[u_i^p \cdot u_j^q \cdot u_k^r \cdots] < E(u_i^p) \cdot E(u_j^q) \cdot E(u_k^r) \cdots. \tag{3.4}$$

This follows from generalization of the relation

$$\frac{\Gamma(n + 1)\Gamma(p + 1)\Gamma(q + 1)\Gamma(r + 1)}{\Gamma(n + p + q + r + 1)} < \frac{[\Gamma(n + 1)]^3 \Gamma(p + 1)\Gamma(q + 1)\Gamma(r + 1)}{\Gamma(n + p + 1)\Gamma(n + q + 1)\Gamma(n + r + 1)}.$$

The above theorem suggests the possibility of test functions for fitted distributions, relative to a universe with a cdf which, since it is merely conditioned by a sufficient hypothesis for the theorem, may be of the non-parametric type.

A test function of the form

$$Y = \sum_m u_i^p, \qquad p \text{ real and positive} \tag{3.5}$$

might first come to mind. If $p = 1$, compensatory effects of deviations reduce the efficiency of the test function. One is thus led first to consider the test function (3.5) for the case $p = 2$.

4. The moments of the probability distribution of $y_m = \Sigma u_i^2$. We are first concerned with the problem of the determination of the moments of the function

$$y_m = \sum_m u_i^2 \tag{4.1}$$

where i ranges over any particular fixed set of m integers which for simplicity is usually taken as the first m.

One first recalls the fact that the result is independent of *which* m indices have been selected; and that the expected value of any combination of powers is independent of which specific subscripts of u_i are involved.

Since the u_i are correlated, principles of combinatory analysis are involved in determining the moments of y_m. One possible way of obtaining the moments is as follows:

Let v_r denote the rth moment of y_m about $y_m = 0$. Thus

$$E[(y_m)^r] = v_r = E[(\sum_m u_i^2)^r]. \tag{4.2}$$

Now in the expansion of $(\sum_m u_i^2)^r$, the sum of the power indices of each term is $2r$. Thus referring back to (3.1) and (3.3) it will be noted that the expected value of each such term will have the common factor

$$1/[n + 2r]_{2r}.$$

Consider a general term of the expansion of $(\sum_m u_i^2)^r$

$$C_{r_1 r_2 \cdots r_k} \cdot u_{i_1}^{2r_1} u_{i_2}^{r_2} \cdots u_{i_k}^{2r_k}, \qquad \text{with} \quad r_1 + r_2 + \cdots + r_k = r.$$

Clearly

$$E(u_{i_1}^{2r_1} u_{i_2}^{2r_2} \cdots u_{i_k}^{2r_k}) = 2r_1 ! \, 2r_2 ! \cdots 2r_k ! / [n + 2r]_{2r}.$$

and the coefficient $C_{r_1 r_2 \cdots r_k}$ is the multinomial coefficient

$$C_{r_1 r_2 \cdots r_k} = \frac{r!}{r_1! r_2! \cdots r_k!}.$$

Now in the expansion of $(\sum_m u_i^2)^r$ group the terms which have the same set of k values of r_i, irrespective of which indices of u_i are involved. The number of such terms (since each involves k different indices) is $\binom{m}{k}$. If $r_1, r_2, \cdots, r_k$,

are all different each combination could be taken in $k!$ different ways. Thus with r's all different and fixed, the sum of all coefficients of terms with same combination of $2r_i$ powers (irrespective of variation of indices of the u_i) is

$$\binom{m}{k} k! \frac{r!}{r_1! r_2! \cdots r_k!}.$$

This would then constitute the total multiplier for

$$2r_1 ! \, 2r_2 ! \cdots 2r_k !/[n + 2r]_{2r}$$

for a given set of k r's which are all different.

If some of r's are repeated, let $k_1, k_2, \cdots, k_s$ denote the number of repetitions of each different r_i ($k_i \geqq 1$, and $k_1 + k_2 + \cdots + k_s = k$). Then each combination of the k r's corresponding to a set of k products could be taken in

$$k!/(k_1! \, k_2! \cdots k_s!)$$

different ways. Hence the lemma:

LEMMA 1. *Consider all admissible sets of k different subscripts of u_i and a fixed set of values of $r = r_1, r_2, \cdots, r_k$ where*

$$r_1 + r_2 + \cdots + r_k = r$$

such that s of these r's are different, and the number of repetitions in the set of r's is given by $k_1 \, k_2 \cdots k_s$ ($k_i \geqq 1$, and $k_1 + k_2 + \cdots + k_s = k$). The composite coefficient of the terms in v_r involving the factor

$$2r_1! \, 2r_2! \cdots 2r_k!/[n + 2r]_{2r}$$

is given by

$$\binom{m}{k} \frac{k!}{k_1! k_2! \cdots k_s!} \cdot \frac{r!}{r_1! r_2! \cdots r_k!}. \tag{4.3}$$

Examples of computation of v_r by means of the above lemma. The first order moment is given by

$$v_1 = E(\sum_m u_i^2) = m \, 2!/[n + 2]_2 . \tag{4.4}$$

The second order moment is given by

$$v_2 = E[(\sum_m u_i^2)^2] = C_1 E(u_i^4) + C_2 E(u_i^2 u_j^2),$$

and determining the values of C_i from Lemma 1,

$$v_2 = \left[m4! + \binom{m}{2} \binom{2!}{2!} \frac{2!}{1!1!} 2!2! \right] \Big/ [n + 4]_4$$

or

$$v_2 = \left[m4! + 8 \binom{m}{2} \right] \Big/ [n + 4]_4 = \left[m + \binom{m}{2} \frac{1}{3} \right] \Big/ \binom{n + 4}{4}. \tag{4.5}$$

Again for the third order moment,

$$v_3 = E[(\sum_m u_i^2)^3] = C_1E(u_i^6) + C_2E(u_i^2u_j^4) + C_3E(u_i^2u_j^2u_k^2),$$

and using Lemma 1,

$$= \left[m6! + \binom{m}{2}\frac{2!}{1!1!}\frac{3!}{1!2!}2!4! + \binom{m}{3}\frac{3!}{3!}\frac{3!}{1!1!1!}2!2!2!\right] \Big/ [n+6]_6$$

$$= \left[m6! + \binom{m}{2}2!3!4! + \binom{m}{3}2!2!2!3!\right] \Big/ [n+6]_6$$

or

$$v_3 = \left[m + \binom{m}{2}\frac{2}{5} + \binom{m}{3}\frac{1}{15}\right] \Big/ \binom{n+6}{6}. \tag{4.6}$$

Similarly writing the fourth moment in the form

$$v_4 = C_1E(u_i^8) + C_2E(u_i^6u_j^2) + C_3E(u_i^4u_j^4) + C_4E(u_i^2u_j^2u_k^4) + C_5E(u_i^2u_j^2u_k^2u_s^2)$$

and using Lemma 1 it reduces to

$$v_4 = \left[m + \binom{m}{2}\frac{2}{7} + \binom{m}{2}\frac{3}{35} + \binom{m}{3}\frac{3}{35} + \binom{m}{4}\frac{1}{105}\right] \Big/ \binom{n+8}{8}. \tag{4.7}$$

Higher order moments of the probability distribution function may be computed as desired.

An alternate method of computing the moments of the distribution of this test function is the following:

Consider a function $g_0(x)$ such that

$$\frac{d^r g_0(0)}{dx^r} = (2r)!, \qquad g_0(0) = 1. \tag{4.8}$$

Thus

$$E[u^{2r}] = [d^r g_0(0)/dx^r]/[n+2r]_{2r}. \tag{4.9}$$

From the principles of combinatory analysis of linear operators, it follows that[1]

$$E[(\sum_m u_i^2)^r] = \frac{d^r[g_0(x)]^m}{dx^r}\Big|_{x=0} / [n+2r]_{2r}. \tag{4.10}$$

Although this is an enlightening analytical form, actual computations seem to be simpler with the use of Lemma 1.

---

[1] One way of seeing this is to first think of the u_i as statistically independent. The numerators of the resulting terms would be the same as in (4.10). When the u_i are taken as dependent, by virtue of (3.1) the numerators will remain the same while all denominators will reduce to $[n+2r]_{2r}$.

Moment generating function. The moment generating function of the probability distribution of y_m can be written as

$$(4.11)\quad E(e^{ty}) = G_0(t, m) = 1 + \sum_{r=1}^{\infty} [d^r(g_0(x))^m/dx^r \mid_{x=0}]/[n + 2r]_{2r}\, t^r/r!$$

with

$$g_0(x) = 1 + 2\,!\,x + 4\,!\,x^2/2! + 6\,!\,x^3/3! + \cdots + (2r)\,!\,x^r/r! + \cdots$$

$$[n + 2r]_{2r} = (n + 2r)(n + 2r - 1) \cdots (n + 1).$$

Although $g_0(x)$ exists only as a formal power series, $G_0(t, m)$ is defined by (4.11) as a power series with positive coefficients, converging for all t.

5. Some comments on test function, $p = 2$. At the present time the study of the test function for $p = 2$ has not gone far enough to justify publication of results. One difficulty is that although its asymptotic distribution function appears to be normal, the convergence towards normalcy may be extremely slow in some cases.

Furthermore there are indications that the case $m = n + 1$ will give the most definitive results not only because the complete range of data is used, but also because errors of Type II would in general have a less erratic effect.

For the case $m = n + 1$ the mean, variance and third and fourth reduced moments (i.e. moments about the mean divided by corresponding power of σ) are:

Case $m = n + 1$.

$$E(y_{n+1}) = 2/(n + 2), \qquad \sigma^2 = 4n/[(n + 2)^2(n + 3)(n + 4)],$$

$$(5.1)\quad \alpha_3 = \mu_3/\sigma^3 = \frac{10n - 4}{(n + 5)(n + 6)} \sqrt{\frac{(n + 3)(n + 4)}{n}}$$

$$\alpha_4 = \left[\frac{n^5 + 101n^2 + 14n - 8}{(n + 5)(n + 6)(n + 7)(n + 8)}\right]\left[\frac{3(n + 3)(n + 4)}{n}\right]$$

$$\alpha_4 - 3 = \frac{6(41n^4 + 241n^3 + 118n^2 - 784n - 48)}{n(n + 5)(n + 6)(n + 7)(n + 8)}.$$

If data is not grouped the test may be applied as follows: Given a function $Q(X)$ which has been fitted to the cdf $F(X)$. From a random sample of size n with x_i ordered as in (2.1) compute the successive differences of $Q(x_i)$ to obtain the variables u_i^*. Then consider the sum of the squares

$$U^* = \sum_{n+1} u_i^{*2}.$$

If $Q(X)$ is a true representation of $F(X)$ the variation of U^* will follow that of y_{n+1}. Thus the expected value of U^*, its variance etc. will be *independent* of the fitted function $Q(X)$, which represents certain advantages over the χ^2 test.

The effect of Type II errors can be roughly analyzed as follows: In considering the effect of such errors the testing procedure must be criticized from the point of view that

$$Q(X) \neq F(X).$$

For $m = n + 1$ it still is true that

$$\Sigma u_i^* = 1$$

which tends to act as a control upon U^*. For example set

$$u_i^* = u_i + \chi_i .$$

Then from the above relation it follows that

$$\Sigma\chi_i = 0. \tag{5.2}$$

Write U^* as

$$\begin{aligned} U^* &= \Sigma u_i^2 + \Sigma\chi_i^2 + 2\Sigma u_i\chi_i \\ &= \Sigma u_i^2 + \Sigma\chi_i^2 + (2\Sigma\chi_i)/(n+1) + 2\Sigma\chi_i\delta(u_i) \end{aligned} \tag{5.3}$$

where $\delta(u_i)$ denotes the variation of the true frequency differences from their expected value $1/(n + 1)$.

The variation $\delta(u_i)$ will be to a considerable degree independent of χ_i. Thus the term $\Sigma\chi_i^2$ will in general tend to be larger than the last term on the right. The third term on the right will be zero by virtue of (5.2), and hence U^* will tend to be larger than y_{n+1}. A similar effect upon the sampling variance of U^* can be noted. Hence an interval of rejection

$$U^* \geqq A, \qquad P[y_{n+1} \leqq A] = \alpha = \text{confidence level},$$

is pointed to.

On the other hand if $m < n + 1$ the condition (5.2) no longer holds, the term $(2\,\Sigma\chi_i)/(n + 1)$ of (5.3) will not be zero and in many cases would dominate the other two error terms. Thus it is easily conceivable that one may have in the case $m < n + 1$

$$U_m^* < y_m$$

even when the discrepancies χ_i are large. Hence in the case $m < n + 1$ choice of confidence interval will require considerable care (see [1]).

Although the distribution of y_{n+1} for small n is decidedly non-normal, if the test function is replaced by

$$r_{n+1} = (\Sigma[u_i - 1/(n+1)]^2)^{\frac{1}{2}} \tag{5.4}$$

it will be found that the probability density function takes on the normal character quite rapidly with increasing n. Indeed the author has found that a computed approximation to the probability density function of r_{n+1} with $n = 4$ is decidedly normal in character.

REFERENCES

[1] J. NEYMAN, "Smooth test for goodness of fit," *Skand. Aktuar. Tidskn.* (1937) p. 149.

[2] E. S. PEARSON, "The probability integral transformation for testing goodness of fit and combining independent tests of significance," *Biometrika*, Vol. 30 (1938), pp. 134–148.

[3] E. J. GUMBEL, "Simple tests for given hypothesis," *Biometrika*, Vol. 32 (1942), pp. 317–333.

[4] H. SCHEFFÉ AND J. W. TUKEY, "Non-parametric estimation, I. Validation of order statistics," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 187–192.

AN ESSENTIALLY COMPLETE CLASS OF ADMISSIBLE DECISION FUNCTIONS

By Abraham Wald

Columbia University

Summary. With any statistical decision procedure (function) there will be associated a risk function $r(\theta)$ where $r(\theta)$ denotes the risk due to possible wrong decisions when θ is the true parameter point. If an a priori probability distribution of θ is given, a decision procedure which minimizes the expected value of $r(\theta)$ is called the Bayes solution of the problem. The main result in this note may be stated as follows: Consider the class C of decision procedures consisting of all Bayes solutions corresponding to all possible a priori distributions of θ. Under some weak conditions, for any decision procedure T not in C there exists a decision procedure T^* in C such that $r^*(\theta) \leqq r(\theta)$ identically in θ. Here $r(\theta)$ is the risk function associated with T, and $r^*(\theta)$ is the risk function associated with T^*. Applications of this result to the problem of testing a hypothesis are made.

1. Introduction. In some previous publications [1], [2] the author has considered the following general problem of statistical inference: Let $X = (X_1, \cdots, X_n)$ be a set of chance variables. Suppose that the only information we have concerning the joint distribution function F of these chance variables is that F is an element of a given class Ω of distribution functions. Suppose, furthermore, that a class D of possible decisions d is given one of which is to be made on the basis of an observation $x = (x_1, \cdots, x_n)$ on the chance vector X. The problem is then to construct a function $d(x)$, called statistical decision function, which associates with each sample point x an element $d(x)$ of D so that the decision $d(x)$ is made when the sample point x is observed. A statistical decision function $d(x)$ is defined over all possible points x of the sample space and for each sample point x the value of the function is an element of D. Each element d of D will usually be interpreted as a decision to accept the hypothesis that the unknown distribution F of X belongs to a certain subclass ω of Ω. Different elements d of D correspond to different subclasses ω of Ω.

The problem of testing the hypothesis H that the unknown distribution function F belongs to a given subclass ω of Ω, is contained as a special case in the above general problem. The space D will then contain only two elements, d_1 and d_2, where d_1 denotes the decision of accepting H and d_2 denotes the decision of rejecting H.

As in [1] and [2], we shall assume also here that Ω is a k-parameter family of distribution functions. Then each element of Ω may be represented by a point $\theta = (\theta_1, \cdots, \theta_k)$, called parameter point, in the k-dimensional Cartesian space. The class Ω is then represented by a subset of the k-dimensional Cartesian space,

called parameter space. We shall, therefore, refer to Ω as the parameter space and to its elements as parameter points.

The merits of any particular decision function $d(x)$ will usually depend on the relative importance of the various possible errors caused by not selecting the proper element d of D. The relative importance of such errors has been described in [1] and [2] by a weight function $W(\theta, d)$ defined over the product of Ω and D. For any pair (θ, d) the value of $W(\theta, d)$ is non-negative and expresses the loss caused by taking the decision d when θ is the true parameter point. For any given decision function $d(x)$ the expected value of the loss is given by

$$r(\theta) = \int_M W[\theta, d(x)]\, dF(x) \tag{1.1}$$

where M denotes the sample space and $F(x)$ is the joint cumulative distribution of $X = (X_1, \cdots, X_n)$ corresponding to the parameter point θ.

The function $r(\theta)$ is defined over the parameter space Ω and is called the risk function. The shape of the risk function $r(\theta)$ will, in general, be affected by the decision function $d(x)$ used. To put this dependence in evidence, we shall use the symbol $r[\theta \mid d(x)]$ to denote the risk function $r(\theta)$ associated with the decision function $d(x)$.

A decision function $d(x)$ is said to be uniformly better than the decision function $d^*(x)$ if

$$r[\theta \mid d(x)] \leqq r[\theta \mid d^*(x)] \tag{1.2}$$

for all θ and if there exists at least one point θ for which the inequality sign holds in (1.2). A decision function $d(x)$ is said to be admissible if no other uniformly better decision function exists.

A class C of admissible decision functions will be said to be essentially complete if for any decision function $d(x)$ not in C there exists a decision function $d^*(x)$ in C such that

$$r[\theta \mid d^*(x)] \leqq r[\theta \mid d(x)]$$

for all θ.

In section 2 we shall formulate certain assumptions which will then be used in section 3 to derive an essentially complete class of admissible decision functions. In section 4 applications are made to the problem of testing a hypothesis.

In a recent paper Lehmann [3] obtained an essentially complete class of admissible tests for each hypothesis H of a certain restricted class of simple hypotheses. The restrictions imposed on Ω in Lehmann's paper are essentially those formulated by Neyman [4], [5] to insure the existence of the type A_1 (uniformly most powerful unbiassed) test. Our definition of an essentially complete class of admissible decision functions agrees with that given by Lehmann when the problem is to test a hypothesis and the weight function $W(\theta, d)$ can take only the values 0 and 1.

2. Assumptions. Throughout this paper we shall make the following assumptions:

Assumption 1: The parameter space Ω is a bounded and closed subset of a finite dimensional, say k-dimensional, Cartesian space.

We shall introduce the following convergence definition in the space D: a sequence $\{d_m\}$, $(m = 1, 2, \cdots, \text{ad inf.})$, of elements of D is said to converge to the element d of D if

$$\lim_{m=\infty} W(\theta, d_m) = W(\theta, d)$$

uniformly in θ.

Assumption 2: The space D is compact and, for any d, $W(\theta, d)$ is a continuous function of θ.

Assumption 3: For any point θ of Ω the joint distribution function of $X = (X_1, \cdots, X_n)$ admits a density function $p(x, \theta)$ for all points x of the n-dimensional Cartesian space M (sample space). The density function $p(x, \theta)$ is assumed to be continuous in x and θ jointly.

In what follows we shall mean by a distribution function $f(\theta)$ of θ a cumulative distribution function for which $\int_\Omega df(\theta) = 1$ and for which $\int_\Omega W(\theta, d) df(\theta)$ is not zero identically in d.

Assumption 4: For any point x of M, except perhaps for a set of measure zero, and for any cumulative distribution function $f(\theta)$ there exists one and only one element of D for which the expression

$$\int_\Omega W(\theta, d) p(x, \theta)\, df(\theta) \tag{2.1}$$

takes its minimum value with respect to d.

Assumptions 1 and 3 in this paper are exactly the same as Assumptions 1 and 3 in [2]. The formulation of Assumptions 2 and 4 is somewhat different from that given in [2]. This is mainly due to the fact that in [2] the space D has the same elements as Ω, while here this is not necessarily so. It can be verified without difficulty that this slight modification of the assumptions does not affect in any way the validity of the results obtained in [2]. Thus, we shall be able to make use of any theorems proved in [2] for the purposes of the present paper.

3. Derivation of an essentially complete class of admissible decision functions. For any distribution function $f(\theta)$ defined over Ω and for any sample point x let $d(x, f)$ denote the element of D for which the expression (2.1) takes its minimum value. It follows easily from the definition of $r(\theta)$ and $d(x, f)$ that

$$\int_\Omega r[\theta \mid d(x, f)] df(\theta) \leqq \int_\Omega r[\theta \mid d^*(x)]\, df(\theta) \tag{3.1}$$

for any decision function $d^*(x)$. If we interpret $f(\theta)$ as an a priori probability distribution of θ, inequality (3.1) says that the expected value of $r(\theta)$ takes its minimum value for the decision function $d(x, f)$. We shall refer to $d(x, f)$ as the Bayes' solution of the problem corresponding to the a priori probability distribution $f(\theta)$.

We shall now prove the following theorem.

THEOREM 3.1. *The class C of all Bayes' solutions $d(x, f)$ corresponding to all possible a priori distributions $f(\theta)$ is an essentially complete class of admissible decision functions.*

PROOF. First we show that for any distribution $f(\theta)$ the decision function $d(x, f)$ is admissible. Let $d(x)$ be a decision function such that

$$r[\theta \mid d(x)] \leqq r[\theta \mid d(x, f)]$$

for all θ. Then

$$\int_\Omega r[\theta \mid d(x)]\, df(\theta) \leqq \int_\Omega r[\theta \mid d(x, f)]\, df(\theta). \tag{3.2}$$

From the definition of $d(x, f)$ it follows that the equality sign must hold in (3.2), i.e.,

$$\int_\Omega r[\theta \mid d(x)]\, df(\theta) = \int_\Omega r[\theta \mid d(x, f)]\, df(\theta). \tag{3.3}$$

From the second half of Theorem 4.2 in [2] it then follows that

$$r[\theta \mid d(x)] = r[\theta \mid d(x, f)]$$

for all θ. Hence $d(x, f)$ is an admissible decision function.

We shall now show that the class C of decision functions $d(x, f)$ corresponding to all possible a priori distributions $f(\theta)$ is essentially complete. Let $d_0(x)$ be any decision function not in the class C. The essential completeness of the class C is proved if we can show that there exists a distribution $f(\theta)$ such that

$$r[\theta \mid d(x, f)] \leqq r[\theta \mid d_0(x)] \tag{3.4}$$

for all θ.

To prove (3.4) we shall consider the weight function

$$W^*(\theta, d) = W(\theta, d) - r[\theta \mid d_0(x)] + \operatorname*{Max}_\theta r[\theta \mid d_0(x)] \tag{3.5}$$

The maximum of $r[\theta \mid d_0(x)]$ exists, since according to Theorem 4.1 in [2] $r[\theta \mid d_0(x)]$ is a continuous function of θ. Clearly, Assumptions 1–4 remain valid if we replace $W(\theta, d)$ by $W^*(\theta, d)$. Let $r^*[\theta \mid d(x)]$ denote the risk function associated with the decision function $d(x)$ if the weight function is given by $W^*(\theta, d)$. According to Theorem 5.2 in [2] there exists a decision function $d^*(x)$ such that

$$\operatorname*{Max}_\theta r^*[\theta \mid d^*(x)] \leqq \operatorname*{Max}_\theta r^*[\theta \mid d(x)] \tag{3.6}$$

for any decision function $d(x)$. Since

$$\operatorname{Max}_{\theta} r^*[\theta \mid d_0(x)] = \operatorname{Max}_{\theta} r[\theta \mid d_0(x)]$$

it follows from (3.6) that

$$\operatorname{Max}_{\theta} r^*[\theta \mid d^*(x)] \leqq \operatorname{Max}_{\theta} r[\theta \mid d_0(x)]. \tag{3.7}$$

Inequalities (3.5) and (3.7) imply

$$r[\theta \mid d^*(x)] \leqq r[\theta \mid d_0(x)] \tag{3.8}$$

for all θ.

For any distribution $f(\theta)$ we shall denote by $d^*(x, f)$ the Bayes solution of the problem corresponding to the a priori distribution $f(\theta)$ when the weight function is given by $W^*(\theta, d)$. Since $W^*(\theta, d) - W(\theta, d)$ depends only on θ but not on d, one can easily verify that $d^*(x, f) = d(x, f)$. It follows from Theorems 4.4 and 5.1 in [2] that there exists a distribution $f(\theta)$, the so-called least favorable distribution, such that (3.6) remains valid if we replace $d^*(x)$ by $d^*(x, f)$. Thus we can put

$$d^*(x) = d^*(x, f) = d(x, f). \tag{3.9}$$

Hence, from (3.8) we obtain

$$r[\theta \mid d(x, f)] \leqq r[\theta \mid d_0(x)]$$

for all θ. This completes the proof of Theorem 3.1.

4. Applications to the problem of testing a hypothesis. In this section we shall apply the results of the preceding section to the problem of testing the hypothesis H that the true parameter point is included in a given subset ω of Ω. We shall assume that ω is an open subset of Ω. The space D consists now only of two elements, d_1 and d_2, where d_1 denotes the decision of accepting H and d_2 denotes the decision of rejecting H.

We shall assume that the $W(\theta, d_1)$ is equal to zero for points θ in the interior or on the boundary of ω, and positive elsewhere. Similarly, $W(\theta, d_2)$ will be assumed to be positive for points θ inside ω and zero outside ω. For any a priori distribution $f(\theta)$ the Bayes solution is given by the following test: We reject the hypothesis H if (and only if)[1]

$$\int_{\Omega-\omega} W(\theta, d_1)p(x, \theta)\, df(\theta) > \int_{\omega} W(\theta, d_2)p(x, \theta)\, df(\theta). \tag{4.1}$$

Thus, the class C of regions (4.1), corresponding to all possible distributions $f(\theta)$, is an essentially complete class of admissible critical regions.

For any critical region R we shall denote the probability that the sample x

[1] Whether the equality sign is included or not in (4.1) is of no consequence, since by Assumption 4 the measure of the set of points x for which the equality holds in (4.1) is zero.

will fall in R when θ is true by $P(\theta \mid R)$. It follows from Lemma 4.4 in [2] and Assumption 3 that $P(\theta \mid R)$ is a continuous function of θ for any region R. Since $W(\theta, d_1)$ is positive in the interior of $\Omega - \omega$, and $W(\theta, d_2)$ is positive in ω, the class C of regions defined in (4.1) will have the following properties:

(a) For any region R outside the class C there exists a region R^* in C such that

$$P(\theta \mid R^*) \leqq P(\theta \mid R) \text{ in } \omega$$

and

$$P(\theta \mid R^*) \geq P(\theta \mid R) \text{ in } \Omega - \omega.$$

(b) If R and R^* are members of C such that

$$P(\theta \mid R^*) \leqq P(\theta \mid R) \text{ in } \omega$$

and

$$P(\theta \mid R^*) \geq P(\theta \mid R) \text{ in } \Omega - \omega,$$

then

$$P(\theta \mid R^*) = P(\theta \mid R) \text{ for all } \theta.$$

For any distribution $g(\theta)$ consider the critical region consisting of all sample points x satisfying

$$\int_{\Omega-\omega} p(x, \theta)\, dg(\theta) > \int_{\omega} p(x, \theta)\, dg(\theta). \tag{4.2}$$

Let C^* be the class of regions (4.2) corresponding to all possible distributions $g(\theta)$. One can easily verify that any region in C is also a member of C^*. Thus, the following theorem holds:

THEOREM 4.1 *Suppose that Assumptions 1 and 3 are fulfilled and ω is an open subset of Ω. Suppose, furthermore, that for any distribution $g(\theta)$ the set of sample points x satisfying the equation*

$$\int_{\Omega-\omega} p(x, \theta)\, dg(\theta) = \int_{\omega} p(x, \theta)\, dg(\theta)$$

has the measure zero. Then, for any region R outside the class C^* there will be a region R^* in C^* such that

$$P(\theta \mid R^*) \leq P(\theta \mid R) \text{ in } \omega$$

and

$$P(\theta \mid R^*) \geqq P(\theta \mid R) \text{ in } \Omega - \omega.$$

Addition at proof reading: After this paper was sent to the printer, the author obtained a generalization of Theorem 3.1 to sequential decision functions, as well as some other results. They will appear in a forthcoming issue of *Econometrica*.

REFERENCES

[1] A. WALD, "Contributions to the theory of statistical estimation and testing hypotheses", *Annals of Math. Stat.*, Vol. 10 (1939).

[2] A. WALD, "Statistical decision functions which minimize the maximum risk", *Annals of Math.*, Vol. 46 (1945).

[3] E. L. LEHMANN, "On families of admissible tests", *Annals of Math. Stat.*, Vol. 18 (1947), pp. 97–104.

[4] J. NEYMAN AND E. S. PEARSON, "Contributions to the theory of testing statistical hypotheses, Part II", *Statistical Research Memoirs*, Vol. II, London, 1938.

[5] J. NEYMAN AND E. S. PEARSON, "Contributions to the theory of testing statistical hypotheses, Part I", *Statistical Research Memoirs*, Vol. I, London, 1936.

DISCRIMINATING BETWEEN BINOMIAL DISTRIBUTIONS

By Paul G. Hoel

University of California at Los Angeles

1. Summary. Given a set of k random samples, $x_1, x_2, \cdots, x_k$, from a binomial distribution with parameters p and n, it is shown that the familiar binomial index of dispersion

$$z = \frac{\sum_1^k (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n_0}\right)}$$

yields an approximate best critical region independent of p for testing the hypothesis $n = n_0$ against the alternative hypothesis $n > n_0$, provided $\bar{x}$ and $n_0 - \bar{x}$ are not small. Because of the nature of the test, its optimum properties also apply to testing whether the data came from a binomial population with $n = n_0$ or from a Poisson population.

2. Introduction. A problem of considerable interest in certain fields is that of deciding whether a set of observations should be treated as having come from either a binomial population or from a Poisson population. Although there was much discussion a few years ago concerning the best method for making such a decision [1], [2], [3], no solution of the problem was presented. In this paper a test that possesses certain optimum properties is derived for discriminating between two binomial populations. This test, however, is also capable of solving the problem of how to discriminate between a binomial and a Poisson population. The methods that are employed in the derivation of this test are similar to those of an earlier paper [4] in which the problem of discriminating between two Poisson populations was studied.

3. Similar regions. Let n denote the number of trials and p the probability of success in a single trial for a binomial distribution. Let $x_1, x_2, \cdots, x_k$ represent the observed frequencies in k random samples from this binomial population. Now consider the two alternative hypotheses

$$H_0 : n = n_0, p = p_0$$

and

$$H_1 : n = n_1 > n_0, p = p_1.$$

The purpose of this paper is to construct a test for discriminating between the two values of n regardless of the values of p; however it is convenient to begin with these more restrictive hypotheses

For the purpose of finding a critical region for testing H_0 against H_1, the x_i will be treated as the coordinates of a point in k dimensions. The probability of obtaining the particular point $x_1, \cdots, x_k$ when H_0 is true will be denoted by $P_0[x_i]$. Since the probability of obtaining x successes in n trials is given by

$$\frac{n!}{x!(n-x)!} p^x q^{n-x}$$

it follows that

$$P_0[x_i] = \frac{(n_0!)^k}{\prod_1^k x_i!(n_0 - x_i)!} p_0^{\sum_1^k x_i} q_0^{\sum_1^k (n_0 - x_i)}. \tag{1}$$

In searching for a critical region that will be independent of p_0, it is illuminating to study the methods that were designed by Neyman and Pearson [5] for continuous distributions. These methods suggest that one should look for critical regions on the surfaces $\sum_1^k x_i = \text{constant}$. For this reason, instead of using (1) for constructing critical regions, it is desirable to study the conditional probability distribution of the points lying in the plane $\sum_1^k x_i = N$, where N is a positive integer not exceeding kn_0. The conditional probability of obtaining the point $x_1, \cdots, x_k$, when the point is restricted to lie in the plane $\sum_1^k x_i = N$, will be denoted by $P_0[x_i \mid N]$. Its value may be obtained by dividing the probability (1) by the probability that the point will lie in the plane $\sum_1^k x_i = N$. If this latter probability is denoted by $P_0[N]$, then

$$P_0[x_i \mid N] = \frac{P_0[x_i]}{P_0[N]}. \tag{2}$$

Since the sum of k independent variables each possessing the same binomial distribution has a binomial distribution with n replaced by kn, it follows that N possesses a binomial distribution and that

$$P_0[N] = \frac{(kn_0)!}{N!(kn_0 - N)!} p_0^N q_0^{kn_0 - N}. \tag{3}$$

If (1) and (3) are substituted in (2), it will reduce to

$$P_0[x_i \mid N] = \frac{(n_0!)^k N!(kn_0 - N)!}{(kn_0)! \prod_1^k x_i!(n_0 - x_i)!}. \tag{4}$$

This conditional probability distribution in the plane $\sum_1^k x_i = N$ is independent of p_0 and therefore may serve as the basis for constructing a critical region that

is independent of p_0 for testing H_0 against H_1. It will therefore be possible to test the less restrictive hypothesis

$$H_0' : n = n_0$$

against

$$H_1' : n = n_1 > n_0 .$$

4. Best critical region. Although a best critical region does not exist for testing H_0' against H_1', it is helpful to proceed as though one did.

If a critical region of size α could be selected in each plane $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$, then the totality of such critical regions would constitute a critical region of size α that is independent of p_0 and which therefore could be used to test H_0' against H_1'. For, if $P_0[X \in \text{C.R.}]$ denotes the probability that the sample point, which will be denoted by X, will lie in the critical region, it follows that

$$\begin{aligned} P_0[X \in \text{C.R.}] &= \sum_{N=0}^{kn_0} P_0[N] P_0[X \in \text{C.R.} \mid N] \\ &= \sum_{N=0}^{kn_0} P_0[N]\alpha \\ &= \alpha. \end{aligned} \tag{5}$$

This last equality follows from the fact that the sample point must lie in one of the planes $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$.

Furthermore, this would be the only critical region of size α independent of p_0, because if a critical region of size α_N, $(N = 0, 1, \cdots, kn_0)$, were selected in the plane $\sum_1^k x_i = N (N = 0, 1, \cdots, kn_0)$, it would be necessary that

$$\sum_{N=0}^{kn_0} P_0[N]\alpha_N = \alpha,$$

independent of the value of p_0. From (3) this is equivalent to requiring that

$$\sum_{N=0}^{kn_0} \frac{(kn_0)!}{N!\,(kn_0 - N)!} p_0^N (1 - p_0)^{kn_0 - N} \alpha_N = \alpha, \tag{6}$$

independent of the value of p_0. Since the left side of (6) is a polynomial in p_0, its constant term must equal α and all other coefficients must vanish. It will be observed that no terms of the sum in (6) that arise from $N > r$ will contribute to the coefficient of p_0^r; consequently this coefficient will not contain the unknowns $\alpha_{r+1}, \cdots, \alpha_{kn_0}$. These considerations show that the α_N must satisfy equations of the form

$$\begin{aligned}
\alpha &= c_{00}\,\alpha_0 \\
0 &= c_{10}\,\alpha_0 + c_{11}\,\alpha_1 \\
&\cdots\cdots \\
0 &= c_{kn_0 0}\,\alpha_0 + c_{kn_0 1}\,\alpha_1 + \cdots + c_{kn_0 kn_0}\,\alpha_{kn_0}.
\end{aligned}$$

It will also be observed that $c_{rr} = (kn_0)!/r!(kn_0 - r)!$; consequently the triangular matrix of the coefficients in these $kn_0 + 1$ non-homogeneous equations is non-singular. The equations therefore possess a unique solution, namely the known solution of $\alpha_N = \alpha$.

The preceding discussion shows that it is necessary to find critical regions of size α in each plane $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$, if a critical region independent of p_0 is desired. If each such planar critical region were a best critical region for that plane, then the totality of such regions would constitute a best critical region independent of p_0 for testing H_0' against H_1'.

It follows from the theory of best critical regions [5] that if a best critical region in the plane $\sum_1^k x_i = N$ did exist, it would be determined by the inequality

$$\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} < K, \tag{7}$$

where P_1 corresponds to P_0 when H_1 is true and where K is a constant whose value is chosen to make the critical region one of size α. Now from (4),

$$\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} = \frac{(n_0!)^k (kn_0 - N)!\,(kn_1)!\,\Pi(n_1 - x_i)!}{(n_1!)^k (kn_1 - N)!\,(kn_0)!\,\Pi(n_0 - x_i)!}. \tag{8}$$

In order to study the possibility of a best critical region, it is therefore necessary to study the possibility of (8) satisfying inequality (7).

5. Approximate best critical region. Unfortunately, because the variables x_i are discrete, it is not possible to find critical regions of exactly size α for arbitrary α as required in (5). Consequently it is necessary to introduce continuous approximating functions for discrete probability functions or to resort to other devices if critical regions of the type discussed in the preceding section are to be obtained.

For the purpose of introducing such approximations, (8) will be written in the following form:

$$\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} = c_1 \frac{(kn_0 - N)!}{\Pi(n_0 - x_i)!}\left(\frac{1}{k}\right)^{kn_0 - N} \div \frac{(kn_1 - N)!}{\Pi(n_1 - x_i)!}\left(\frac{1}{k}\right)^{kn_1 - N}, \tag{9}$$

where c_1 is independent of the variables x_i. It will be observed that the ratio on the right is a ratio of two multinomial functions. Now the multinomial function

$$\frac{N!}{x_1!x_2!\cdots x_k!}\,p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k},$$

where $\sum_1^k x_i = N$, can be approximated by the multivariate normal function

$$\frac{e\exp\left[-\frac{1}{2}\sum_1^k\left(\frac{x_i - Np_i}{\sqrt{Np_i}}\right)^2\right]}{(2\pi N)^{\frac{1}{2}(k-1)}\sqrt{p_1p_2\cdots p_k}}. \tag{10}$$

The approximation is good provided the Np_i are large and the x_i remain away from their extreme values. If this approximation is applied to both numerator and denominator of (9), to this order of approximation,

$$\begin{aligned}\frac{P_0[x_i \mid N]}{P_1[x_i \mid N]} &= c_1\,\frac{k^{k/2}e\exp\left[-\frac{1}{2}\sum_1^k\left(\frac{x_i - N/k}{\sqrt{n_0 - N/k}}\right)^2\right]}{[2\pi(kn_0 - N)]^{\frac{1}{2}(k-1)}} \\ &\quad \div \frac{k^{k/2}e\exp\left[-\frac{1}{2}\sum_1^k\left(\frac{x_i - N/k}{\sqrt{n_1 - N/k}}\right)^2\right]}{[2\pi(kn_1 - N)]^{\frac{1}{2}(k-1)}} \\ &= c_1\left[\frac{kn_1 - N}{kn_0 - N}\right]^{\frac{1}{2}(k-1)} e\exp\left[-\frac{1}{2}\,\frac{n_1 - n_0}{(n_1 - N/k)(n_0 - N/k)}\right. \\ &\qquad \left.\cdot\sum_1^k (x_i - N/k)^2\right].\end{aligned} \tag{11}$$

Since, by hypothesis, $n_1 > n_0$ and $n_0 > N/k$, except for the case of $n_0 = N/k$, which will be considered later, it follows that

$$\frac{n_1 - n_0}{(n_1 - N/k)(n_0 - N/k)} > 0.$$

As a consequence, the right side of (11) will decrease in value as $\sum_1^k (x_i - N/k)^2$ increases in value. If $(x_1, \cdots, x_k)$ is a point lying on the sphere

$$\sum_1^k (x_i - N/k)^2 = R \tag{12}$$

and if the coordinates of this point satisfy inequality (7) when approximation (11) is used, then all points outside this sphere will also satisfy (7) to this same order of approximation. A best critical planar region of size α in this approximate sense can therefore be obtained in the plane $\sum_1^k x_i = N$ by determining a

sphere with center at $(N/k, \cdots, N/k)$ such that when H_0' is true the probability is α that a point lying in the plane will lie outside this sphere. Furthermore, such a region will be a common best critical region for all values of $n_1 > n_0$ because the preceding arguments do not require the value of n_1 but merely the knowledge that $n_1 > n_0$.

For the purpose of determining the radius of the sphere that will yield the desired critical region, (4) will be expressed as follows:

$$P_0[x_i \mid N] = c_2 \frac{N!}{\Pi x_i!}\left(\frac{1}{k}\right)^N \frac{(kn_0 - N)!}{\Pi(n_0 - x_i)!}\left(\frac{1}{k}\right)^{kn_0 - N}, \tag{13}$$

where c_2 is independent of the x_i. If these multinomials are replaced by their multivariate normal approximations as given by (10), to this approximation (13) will reduce to

$$\begin{aligned} P_0[x_i \mid N] &= c_3\, e \exp\left[-\tfrac{1}{2}\sum_1^k \left(\frac{x_i - N/k}{\sqrt{N/k}}\right)^2\right] e \exp\left[-\tfrac{1}{2}\sum_1^k \left(\frac{x_i - N/k}{\sqrt{n_0 - N/k}}\right)^2\right] \\ &= c_3\, e \exp\left[-\tfrac{1}{2}\frac{\sum_1^k (x_i - N/k)^2}{\frac{N}{k}\left(1 - \frac{N}{kn_0}\right)}\right] \end{aligned} \tag{14}$$

where c_3 is independent of the x_i. Since $\sum_1^k x_i = N$ here, x_k may be expressed in terms of the remaining variables; consequently (14), except for a constant factor, may be treated as a normal distribution in the variables $x_1, \cdots, x_{k-1}$. If the factorials in c_3 are replaced by their Stirling approximations, it will be found that c_3 is the correct constant for the normal distribution.

Since it is known [6] that -2 times the exponent in a normal distribution function possesses a chi-square distribution, it follows that to this order of approximation

$$\frac{\sum_1^k (x_i - N/k)^2}{\frac{N}{k}\left(1 - \frac{N}{kn_0}\right)} \tag{15}$$

possesses a chi-square distribution with $k - 1$ degrees of freedom. If χ_α^2 is a value such that $P[\chi^2 > \chi_\alpha^2] = \alpha$, then

$$\frac{\sum_1^k (x_i - N/k)^2}{\frac{N}{k}\left(1 - \frac{N}{kn_0}\right)} = \chi_\alpha^2 \tag{16}$$

determines a sphere such that to this order of approximation the probability is α that a point lying in the plane $\sum_1^k x_i = N$ will lie outside the sphere. From

the arguments following (12), it therefore follows that a common best critical region in this approximate sense for testing H_0' against H_1' will consist of that part of each plane $\sum_1^k x_i = N$, $(N = 0, 1, \cdots, kn_0)$, which lies outside the corresponding sphere given by (16). Since the x_i are non-negative and do not exceed n_0, the planes corresponding to $N = 0$ and $N = kn_0$ contain a single point; therefore it is necessary to adopt some convention that assigns 100α percent of the samples with $N = 0$ and $N = kn_0$ to a critical region in order to obtain critical regions of size α in these two cases.

For a given set of data, the procedure to be followed then consists in calculating the statistic

$$z = \frac{\sum_1^k (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n_0}\right)},$$

where $\bar{x} = \sum_1^k x_i/k$, and agreeing to reject the hypothesis that $n = n_0$ in favor of the alternative hypothesis that $n > n_0$ if and only if $z > \chi_\alpha^2$, where $P[\chi^2 > \chi_\alpha^2] = \alpha$ for $k - 1$ degrees of freedom. Because of the nature of the approximations used in (10) and (14), this result may be expected to be accurate only if $\bar{x}$ and $n_0 - \bar{x}$ are large.

The interesting feature of this result is that the familiar binomial index of dispersion, z, possesses optimum properties in this approximate sense for testing $n = n_0$ against $n > n_0$.

6. Poisson application. Since the preceding test will possess approximate optimum properties for n as large as desired, independent of the value of p, and since a Poisson distribution with parameter m can be approximated as closely as desired by means of a binomial distribution with $np = m$ by allowing n to increase sufficiently, it follows that the test will also possess approximate optimum properties for deciding between a binomial distribution with $n = n_0$ and a Poisson distribution.

7. Estimation of n. Although the purpose of this paper has been accomplished in the preceding sections, it is interesting to observe the role played by the closely related Poisson index of dispersion in the extimation of n.

Approximate confidence limits for n may be obtained by means of (16). If $\chi_{1-\alpha}^2$ is a value of χ^2 such that $P[\chi^2 > \chi_{1-\alpha}^2] = 1 - \alpha$, then, to this same order of approximation, the probability is $1 - 2\alpha$ that

$$\chi_{1-\alpha}^2 < \frac{\sum_1^k (x_i - \bar{x})^2}{\bar{x}\left(1 - \frac{\bar{x}}{n}\right)} < \chi_\alpha^2.$$

If these inequalities are solved for n, the following $100(1 - 2\alpha)$ percent approximate confidence limits for n will be obtained:

$$(17) \qquad \frac{\bar{x}\chi_{\alpha}^2}{\chi_{\alpha}^2 - \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}} < n < \frac{\bar{x}\chi_{1-\alpha}^2}{\chi_{1-\alpha}^2 - \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}}.$$

Only the lower limit here will possess optimum properties. Now it will be observed that only positive values of n will be admissible if

$$\frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} \leq \chi_{1-\alpha}^2,$$

whereas only negative values will be admissible if

$$\frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} \geq \chi_{\alpha}^2.$$

The range of values will be infinite in each case if there is equality rather than inequality. If, however,

$$\chi_{1-\alpha}^2 < \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}} < \chi_{\alpha}^2,$$

then both positive and negative values of n over infinite ranges will be admissible. Since n increases as the Poisson index $\Sigma(x_i - \bar{x})^2/\bar{x}$ increases until it becomes infinite and then increases from minus infinity through negative values, (17) may still be thought of as giving an interval (infinite) of values with a positive "lower" limit and a negative "upper" limit. Thus, the familiar Poisson index of dispersion plays an interesting role in determining whether a Poisson assumption is reasonable as far as admissible values of n are concerned.

If the population is truly binomial, negative values of n must be ruled out; consequently a Poisson assumption becomes increasingly tenable as the Poisson index increases. However, experience has shown [7] that a negative binomial distribution is often more realistic in describing data supposedly drawn from a binomial or Poisson population than is the assumed distribution; consequently a negative binomial should be given consideration if (17) yields only negative values or if it yields a negative "upper" limit that is numerically small relative to a positive "lower" limit.

It is also interesting to consider the point estimation of n. Here, it is customary [7] to estimate n by means of

$$\frac{k\bar{x}}{k - \frac{\Sigma(x_i - \bar{x})^2}{\bar{x}}}.$$

Thus, a positive, infinite, or negative estimate for n will be obtained according as the Poisson index is less than, equal to, or greater than k.

REFERENCES

[1] J. Berkson, "Some difficulties of interpretation encountered in the application of the chi-square test," *Jour. Amer. Stat. Assoc.*, Vol. 33 (1938), pp. 526–536.

[2] B. H. Camp, "Further interpretations of the chi-square test," *Jour. Amer. Stat. Assoc.*, Vol. 33 (1938), pp. 537–542.

[3] J. Berkson, "A note on the chi-square test, the Poisson, and the binomial," *Jour. Amer. Stat. Assoc.*, Vol. 35 (1940), pp. 362–367.

[4] P. G. Hoel, "Testing the homogeneity of Poisson frequencies," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 362–368.

[5] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Roy. Soc. Phil. Trans.*, Vol. 231 (1933), pp. 289–337.

[6] S. S. Wilks, *Mathematical Statistics*, Princeton Univ. Press, 1943, p. 104.

[7] "Student," "An explanation of deviations from Poisson's law in practice," *Biometrika*, Vol. 12 (1919), pp. 211–215.

BILINEAR FORMS IN NORMALLY CORRELATED VARIABLES

By Allen T. Craig

University of Iowa

1. Summary. If a variable x is normally distributed with mean zero, we have previously given a necessary and sufficient condition (see references at end of this paper) for the independence of two real symmetric quadratic forms in n independent values of that variable. This condition is that the product of the matrices of the forms should vanish. In the present paper, we have proved that the same algebraic condition is both necessary and sufficient for the independence of two real symmetric bilinear, or a real symmetric bilinear and quadratic form, in normally correlated variables.

2. Introduction. In this paper, we determine the moment generating function of the joint distribution of two real symmetric bilinear forms in certain normally correlated variables and derive a necessary and sufficient condition for the independence, in the probability sense, of these forms. We further investigate the condition for independence, in the probability sense, of real symmetric bilinear and quadratic forms.

3. The moment generating function of the distribution of real symmetric bilinear forms. Let the two variables x and y have a joint normal distribution with means zero, unit variances and correlation coefficient ρ. From this bivariate distribution, repeated random samples of n pairs, say (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n), are drawn. Let $C = \| c_{jk} \|$ be a real symmetric matrix and write $\theta = \Sigma\Sigma\, c_{jk}x_jy_k$. The moment generating function of the distribution of θ is then given by

$$\varphi(t) = E[e^{t\theta}] = \frac{1}{(2\pi\sqrt{1-\rho^2})^n}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} e^{t\theta - Q}\, dy_n\, dx_n \cdots dy_1\, dx_1,$$

where

$$Q = \frac{1}{2(1-\rho^2)}\sum_j (x_j^2 + y_j^2 - 2\rho x_j y_j)$$

and θ is defined above. If we subject the x's and y's to the same linear homogeneous transformation with appropriately chosen orthogonal matrix L, then Q remains invariant and θ becomes $\Sigma_j \lambda_j x_j' y_j'$ where the λ's are the n real roots of the characteristic equation of C, that is, of $|\, C - \lambda I\,| = 0$. The integrations are then easily effected and we find that

$$\begin{aligned}\varphi(t) &= \{\prod_j [1 - t(\rho+1)\lambda_j][1 - t(\rho-1)\lambda_j]\}^{-\frac{1}{2}},\\ &= \{\,|\, I - t(\rho+1)C\,| \cdot |\, I - t(\rho-1)C\,|\,\}^{-\frac{1}{2}},\\ &= |\, I - 2\rho tC - (1-\rho^2)t^2C^2\,|^{-\frac{1}{2}},\end{aligned}$$

where I is the unit matrix of order n and the vertical bars, as usual, indicate the determinant of the enclosed matrix.

Next, let $A = \| a_{jk} \|$ and $B = \| b_{jk} \|$ be two real symmetric matrices each of order n. Write $\theta_1 = \Sigma\Sigma a_{jk}x_jy_k$ and $\theta_2 = \Sigma\Sigma b_{jk}x_jy_k$ where the x's and y's are the items of the sample randomly drawn from the bivariate distribution previously described. The moment generating function of the joint distribution of θ_1 and θ_2 is then given by

$$\begin{aligned}\varphi(t_1, t_2) &= E[e^{t_1\theta_1+t_2\theta_2}] \\ &= (2\pi\sqrt{1-\rho^2})^{-n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1\theta_1+t_2\theta_2-Q}\, dy_n\, dx_n \cdots dy_1\, dx_1,\end{aligned}$$

where θ_1, θ_2, and Q have the meanings previously assigned to them. If we pursue a line of reasoning similar to that above, we find that

$$\varphi(t_1, t_2) = | I - 2\rho(t_1A + t_2B) - (1-\rho^2)(t_1A + t_2B)^2 |^{-\frac{1}{2}}.$$

4. The independence of bilinear forms. It is clear that there exist positive numbers, say h_1 and h_2, such that $\varphi(t_1, t_2)$ exists for $0 < t_1 < h_1$ and $0 < t_2 < h_2$. It is well known that a necessary and sufficient condition for the independence of θ_1 and θ_2 is that $\varphi(t_1, t_2)$ shall factor into the product $\varphi(t_1, 0)\varphi(0, t_2)$. If then, we assume θ_1 and θ_2 to be independent, we have essentially

$$\begin{aligned}&| I - 2\rho(t_1A + t_2B) - (1-\rho^2)(t_1A + t_2B)^2 | \\ &\quad = | I - 2\rho t_1A - (1-\rho^2)t_1^2A^2 | \cdot | I - 2\rho t_2B - (1-\rho^2)t_2^2B^2 |.\end{aligned} \tag{1}$$

If h denotes the smaller of h_1 and h_2, then the factored form holds for $0 < t_1, t_2 < h$, and hence for all real values of t_1 and t_2. In particular it holds for $t_2 = t_1$ so that

$$\begin{aligned}&| I - 2\rho t_1(A + B) - (1-\rho^2)t_1^2(A + B)^2 | \\ &\quad = | I - 2\rho t_1A - (1-\rho^2)t_1^2A^2 | \cdot | I - 2\rho t_1B - (1-\rho^2)t_1^2B^2 |.\end{aligned}$$

Let r_1, r_2, and $r \leq r_1 + r_2$ denote the ranks of the matrices A, B, and $A + B$. Further let the real non-zero roots of the characteristic equations of these matrices be denoted respectively by $\alpha_1, \alpha_2, \cdots, \alpha_{r_1}, \beta_1, \beta_2, \cdots, \beta_{r_2}$, and $\gamma_1, \gamma_2, \cdots, \gamma_r$. Then the members of the preceding equation may be written

$$\prod_{i=1}^{r} [1 - t_1(\rho + 1)\gamma_i][1 - t_1(\rho - 1)\gamma_i]$$

and

$$\prod_{i=1}^{r_1} [1 - t_1(\rho + 1)\alpha_i][1 - t_1(\rho - 1)\alpha_i] \prod_{i=1}^{r_2} [1 - t_1(\rho + 1)\beta_i][1 - t_1(\rho - 1)\beta_i]$$

respectively. It is seen that the left member is a polynomial in t_1 of degree $2r$ and that the right member is a polynomial in t_1 of degree $2(r_1 + r_2)$. Accord-

ingly, $r = r_1 + r_2$ and the roots $\gamma_1, \cdots, \gamma_1$ consist of the roots $\alpha_1, \cdots, \alpha_{r_1}$, $\beta_1, \cdots, \beta_{r_2}$. That is, if θ_1 and θ_2 are independent, then the rank of $A + B$ is the sum of the ranks of A and B and the non-zero roots of the characteristic equation of $A + B$ consist of those of the characteristic equation of A together with those of B. Further, if in (1) we put $t_2 = vt_1$, where v is real, we have

$$\begin{aligned}&|I - 2\rho t_1(A + vB) - (1 - \rho^2)t_1^2(A + vB)^2|\\ &\quad = |I - 2\rho t_1 A - (1 - \rho^2)t_1^2A^2| \cdot |I - 2\rho t_1 vB - (1 - \rho^2)t_1^2v^2B^2|.\end{aligned}$$

Denote the rank of $A + vB$ by r' and the non-zero roots of its characteristic equation by $\delta_1, \cdots, \delta_{r'}$. The immediately preceding equation can then be written

$$\prod_{i=1}^{r'} [1 - t_1(\rho + 1)\delta_i][1 - t_1(\rho - 1)\delta_i]$$

$$= \prod_{i=1}^{r_1} [1 - t_1(\rho + 1)\alpha_i][1 - t_1(\rho - 1)\alpha_i] \prod_{i=1}^{r_2} [1 - t_1(\rho + 1)v\beta_i][1 - t_1(\rho - 1)v\beta_i].$$

From this we infer that, apart from zero roots, the roots of the characteristic equation of $A + vB$ are $\alpha_1, \cdots, \alpha_{r_1}, v\beta_1, \cdots, v\beta_{r_2}$.

If a symmetric matrix, say $M(v)$, has elements which are real polynomials in the real variable v, and if the determinant

$$|M(v) - \lambda I| = (-1)^n[\lambda - p_1(v)][\lambda - p_2(v)] \cdots [\lambda - p_n(v)],$$

where $p_1(v), p_2(v), \cdots, p_n(v)$ are likewise real polynomials in v, then there exists, for all real values of v, a real orthogonal matrix, say $L(v)$, such that

$$L'(v)M(v)L(v) = \begin{Vmatrix} p_1(v) & 0 & \cdots & 0 \\ 0 & p_2(v) & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ 0 & & & p_n(v) \end{Vmatrix}.$$

Furthermore[1], $\dfrac{dL(v)}{dv}$ exists for all real values of v. Since

$$|A + vB - \lambda I| = (-1)^n\lambda^{n-(r_1+r_2)}(\lambda - \alpha_1) \cdots (\lambda - \alpha_{r_1})(\lambda - v\beta_1) \cdots (\lambda - v\beta_{r_2}),$$

[1] A number of years ago, in connection with another problem, the writer sought the assistance of Professor N. H. McCoy for a proof that $L(v)$ is differentiable at $v = 0$. Professor McCoy's elegant demonstration of the existence of $L(v)$ showed that each element of this orthogonal matrix is itself a real polynomial in v, divided by the positive square root of another real polynomial, which polynomial is never negative and which vanishes for no real value of v. Thus the derivative of $L(v)$ exists not only for $v = 0$ but for all real values of v. The writer thanks Professor McCoy for his kind and generous assistance.

then $A + vB$ belongs to the class $M(v)$ so we have

$$(2) \qquad L'(v)(A + vB)L(v) = \begin{Vmatrix} \alpha_1 & 0 & \cdots & & 0 \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ 0 & \cdots & \alpha_{r_1} & \cdots & 0 \\ 0 & \cdots & v\beta_1 & \cdots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 0 & \cdots & v\beta_{r_2} & \cdots & 0 \\ \cdot & & & \cdot & \\ \cdot & & & \cdot & \\ \cdot & & & \cdot & \\ 0 & & & & 0 \end{Vmatrix}.$$

In particular,

$$(3) \qquad L'(0)AL(0) = \begin{Vmatrix} \alpha_1 & \cdots & 0 \\ \cdot & & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ & \alpha_{r_1} & \\ & 0 & \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 0 & & 0 \end{Vmatrix}.$$

If we differentiate (2) with respect to v and subsequently set $v = 0$, we have

$$(4) \qquad \frac{dL'(0)}{dv} AL(0) + L'(0)BL(0) + L'(0)A\frac{dL(0)}{dv} = \begin{Vmatrix} 0 & \cdots & 0 \\ \cdot & & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ & 0 & \\ & \beta_1 & \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ & \beta_{r_2} & \\ & \cdot & \\ & \cdot & \\ & \cdot & \\ 0 & & 0 \end{Vmatrix}.$$

Since $L(v)$ is orthogonal, then $L'(v)L(v) = I$. Upon differentiating both members with respect to v, and subsequently setting $v = 0$, it is seen that $\frac{dL'(0)}{dv} L(0) = -L'(0) \frac{dL(0)}{dv}$ so that $L'(0) \frac{dL(0)}{dv}$ is a skew-symmetric matrix, say S. Further

$$\frac{dL'(0)}{dv} = -L'(0) \frac{dL(0)}{dv} L'(0) = -SL'(0), \tag{5}$$

and, by taking conjugates,

$$\frac{dL(0)}{dv} = -L(0) \frac{dL'(0)}{dv} L(0) = L(0)S. \tag{6}$$

If we multiply (5) on the right by AL(0) and (6) on the left by $L'(0)A$, we see that (4) may be written

$$L'(0)BL(0) = \begin{Vmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ & 0 & \\ & \beta_1 & \\ & & \\ & & \\ & \beta_{r_2} & \\ & 0 & \\ \vdots & & \vdots \\ 0 & & 0 \end{Vmatrix} + SL'(0)AL(0) - L'(0)AL(0)S. \tag{7}$$

Since S is skew-symmetric and since $L'(0)AL(0)$ is given by (3), then each element on the principal diagonal of $SL'(0)AL(0)$ and $L'(0)AL(0)S$ is zero. Further, since $L'(0)BL(0)$ is symmetric, then $L'(0)BL(0)$ takes the form

$$\begin{Vmatrix} 0 & k_{12} & \cdots & k_{1n} \\ k_{12} & 0 & & \\ \vdots & & & \\ & & \beta_1 & \\ & & & \\ & & & \\ & & \beta_{r_2} & \\ & & & \\ k_{1n} & & & 0 \end{Vmatrix}.$$

Because the non-zero roots of the characteristic equation of $L'(0)BL(0)$ are $\beta_1, \cdots, \beta_{r_2}$ then the sum of all two-rowed principal minors of the determinant of $L'(0)BL(0)$ must equal the sum of the products of $\beta_1, \cdots, \beta_{r_2}$ taken two at a time. That is

$$\sum_{i<j} \beta_i \beta_j = \sum_{i<j} \beta_i \beta_j - \Sigma k_{ij}^2,$$

so that each k_{ij}, being real, is zero. Accordingly, $SL'(0)AL(0) - L'(0)AL(0)S$ is a zero matrix and $L'(0)BL(0)$ is given by the first term in the right member of (7). We then have

$$L'(0)AL(0)L'(0)BL(0) = L'(0)ABL(0) = 0,$$

from which it follows that $AB = 0$. Thus, if the real symmetric bilinear forms θ_1 and θ_2 are independent in the probability sense, the product of their matrices is zero.

If, conversely, $AB = 0$, then

$$\begin{aligned}\varphi(t_1, t_2) &= |I - 2\rho(t_1A + t_2B) - (1 - \rho^2)(t_1^2A^2 + t_2^2B^2)|^{-\frac{1}{2}}, \\ &= |[I - 2\rho t_1A - (1 - \rho^2)t_1^2A^2][I - 2\rho t_2B - (1 - \rho^2)t_2^2B^2]|^{-\frac{1}{2}}, \\ &= \varphi(t_1, 0)\varphi(0, t_2),\end{aligned}$$

and θ_1 and θ_2 are independent. This establishes the following theorem.

THEOREM I. *Let x and y be normally correlated with means zero, unit variances, and correlation coefficient ρ. Let θ_1 and θ_2 be two real symmetric bilinear forms in n random pairs of values of x and y, say $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$. A necessary and sufficient condition that θ_1 and θ_2 be independent in the probability sense is that the product of the matrices of the forms be zero.*

5. Simultaneous reduction of quadratic or bilinear forms. The argument of Section 4 may be used to establish in a very simple manner the following theorem.

THEOREM II. *Let A and B be two real symmetric matrices with constant elements, each matrix of order n. A necessary and sufficient condition that there exist a real orthogonal matrix of order n such that simultaneously each of $L'AL$ and $L'BL$ is in canonical form, wherein no non-zero elements occupy corresponding positions on the principal diagonals, is that $AB = 0$.*

For if such an orthogonal matrix L exists, it is evident that $L'ALL'BL = L'ABL = 0$ from which it follows that $AB = 0$. Conversely, if $AB = 0$, then v being a real scalar, the matrix $(A - \lambda I)(vB - \lambda I)$ is equal to the matrix $-\lambda[(A + vB) - \lambda I]$. These matrices being equal, their determinants are equal so that $A + vB$ belongs to the class $M(v)$ of section 4. Thus L may be taken as $L(0)$ and simultaneously $L'AL$ and $L'BL$ are of the form stated in the theorem.

6. Independence of bilinaer and quadratic forms. Let $\theta = \Sigma\Sigma a_{jk}x_jy_k$ be a real symmetric bilinear form of rank r_1 in the previously defined variables

$(x_1, y_1), \cdots, (x_n, y_n)$ and let $q = \Sigma\Sigma b_{jk}x_jx_k$ be a real symmetric quadratic form of rank r_2 in $x_1, x_2, \cdots, x_n$. As usual, denote the non-zero roots of the characteristic equations of A and B by $\alpha_1, \alpha_2, \cdots, \alpha_{r_1}$ and $\beta_1, \beta_2, \cdots, \beta_{r_2}$ respectively. The moment generating function of the joint distribution of θ and q is

$$\varphi(t_1, t_2) = \frac{1}{(2\pi\sqrt{1-\rho^2})^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{t_1\theta + t_2 q - Q}\, dy_n\, dx_n \cdots dy_1\, dx_1,$$

where, as previously,

$$Q = \frac{1}{2(1-\rho^2)} \Sigma(x_j^2 + y_j^2 - 2\rho x_j y_j).$$

We first orthogonally transform the variables so that the exponent in the integrand becomes, upon writing $\| f_{jk} \| = L'BL$,

$$t_1 \Sigma\alpha_j x_j' y_j' + t_2 \Sigma\Sigma f_{jk} x_j' x_k' - \frac{1}{2(1-\rho^2)} \Sigma(x_j'^2 + y_j'^2 - 2\rho x_j' y_j').$$

We then integrate on $y_1', y_2', \cdots, y_n'$ and obtain for the exponent in the integrand

$$t_2 \Sigma\Sigma f_{jk} x_j' x_k' - \tfrac{1}{2}\Sigma x_j'^2 + \rho t_1 \Sigma\alpha_j x_j'^2 + \frac{1-\rho^2}{2} t_1^2 \Sigma\alpha_j^2 x_j'^2.$$

If we effect on the variables $x_1', x_2', \cdots, x_n'$ the inverse of the orthogonal transformation initially used on the x's and y's, the exponent in the integrand becomes, using $\| g_{jk} \| = A^2$,

$$t_2 \Sigma\Sigma b_{jk} x_j x_k - \tfrac{1}{2}\Sigma x_j^2 + \rho t_1 \Sigma\Sigma a_{jk} x_j x_k + \frac{1-\rho^2}{2} t_1^2 \Sigma\Sigma g_{jk} x_j x_k$$

or

$$-\tfrac{1}{2}\Sigma\Sigma[\delta_{jk} - 2\rho t_1 a_{jk} - (1-\rho^2)t_1^2 g_{jk} - 2t_2 b_{jk}]x_j x_k,$$

where δ_{jk} equals 1 or 0 according as j does or does not equal k. Hence,

$$\varphi(t_1, t_2) = |\, I - 2\rho t_1 A - (1-\rho^2)t_1^2 A^2 - 2t_2 B\,|^{-\frac{1}{2}}. \tag{8}$$

If θ and q are independent, we have

$$|\, I - 2\rho t_1 A - (1-\rho^2)t_1^2 A^2 - 2t_2 B\,| = |\, I - 2\rho t_1 A - (1-\rho^2)t_1^2 A^2\,| \cdot |\, I - 2t_2 B\,|, \tag{9}$$

for $0 < t_1 < h_1$ and $0 < t_2 < h_2$. As before, the members of (9) are polynomials which, being equal for $0 < t_1, t_2 < h$, are equal for all real values of t_1 and t_2. If we put $t_1 = 1$ and $t_2 = vt_1 = v$, where v is real, then (9) becomes

$$|\, I - 2\rho A - (1-\rho^2)A^2 - 2vB\,| = |\, I - 2\rho A - (1-\rho^2)A^2\,| \cdot |\, I - 2vB\,|$$
$$= \prod_1^{r_1} [1 - (\rho - 1)\alpha_j][1 - (\rho + 1)\alpha_j] \prod_1^{r_2} (1 - 2v\beta_j).$$

That is,

$$|\,2\rho A + (1-\rho^2)A^2 + 2vB - \lambda I\,|$$
$$= (-1)^n \lambda^{n-(r_1+r_2)} [\lambda - 2\rho\alpha_1 - (1-\rho^2)\alpha_1^2]\cdots[\lambda - 2\rho\alpha_{r_1} - (1-\rho^2)\alpha_{r_1}^2]$$
$$\cdot[\lambda - 2v\beta_1]\cdots[\lambda - 2v\beta_{r_2}]$$

so that $2\rho A + (1-\rho^2)A^2 + 2vB$ is a matrix of the class $M(v)$. Hence we write

$$L'(v)[2\rho A + (1-\rho^2)A^2 + 2vB]L(v) = \begin{Vmatrix} 2\rho\alpha_1 + (1-\rho^2)\alpha_1^2 & \cdots & 0 \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \cdot & 2\rho\alpha_{r_1} + (1-\rho^2)\alpha_{r_1}^2 & \\ \cdot & 2v\beta_1 & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & 2v\beta_{r_2} & \\ \cdot & 0 & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ \cdot & \cdot & \\ 0 & & 0 \end{Vmatrix}.$$

The argument of section 4 shows that $L'(0)[2\rho A + (1-\rho^2)A^2]L(0)L'(0)2BL(0)$ is a zero matrix, from which it follows that $2\rho AB + (1-\rho^2)A^2B = 0$. But this imposes on ρ, n^2 conditions of the form

$$2\rho l_{jk} + (1-\rho^2)m_{jk} = 0, \qquad (j, k = 1, 2, \cdots, n).$$

Since these hold for every $-1 < \rho < 1$, they hold identically. Hence each l_{jk} and m_{jk} is zero. In particular, $\| l_{jk} \| = AB = 0$ if θ and q are independent.

Conversely, if $AB = 0$, we see by Theorem II that (8) becomes

$$\varphi(t_1, t_2) = \varphi(t_1, 0)\varphi(0, t_2),$$

so that θ and q are independent. This yields Theorem III.

THEOREM III. *Let x and y be normally correlated with means zero, unit variances, and correlation coefficient ρ. Let θ be a real symmetric bilinear form in the n random pairs of values of x and y, say $(x_1, y_1), \cdots, (x_n, y_n)$, and let q be a real symmetric quadratic form in $x_1, x_2, \cdots, x_n$ (or $y_1, \cdots, y_n$). A necessary and sufficient condition that θ and q be independent in the probability sense is that the product of the matrices of the forms be zero.*

For example, let θ be n times the sample covariance and let q be n times the square of the mean of the x's. Then

$$\theta = \Sigma(x_j - \bar{x})(y_j - \bar{y})$$
$$= \Sigma\Sigma a_{jk} x_j y_k,$$

where

$$a_{jk} = \frac{n-1}{n} \quad \text{if} \quad j = k,$$

$$= -\frac{1}{n} \text{ otherwise,}$$

and

$$q = n\bar{x}^2 = \Sigma\Sigma b_{jk}x_jx_k\,, \qquad b_{jk} = 1/n \text{ for } j, k = 1, 2, \cdots, n.$$

Since $AB = 0$, then θ and q are independent, a fact otherwise known but perhaps not so easily established.

REFERENCES

[1] ALLEN T. CRAIG, "Note on the independence of certain quadratic forms", *Annals of Math. Stat.*, Vol. 14 (1943), pp. 195–197.

[2] HAROLD HOTELLING, "Note on a matric theorem of A. T. Craig", *Annals of Math. Stat.*, Vol. 15 (1944), pp. 427–429.

ON THE CHARLIER TYPE B SERIES

By S. Kullback

George Washington University

1. Introduction. The Type B series of Charlier has been discussed in some detail in the literature (See references at the end of the paper). The problem of the convergence of the Type B series has been considered by Pollaczek-Geiringer [12], [13], Szegö [12] (page 110), Uspensky [16], Jacob [5], Schmidt [16] and Obrechkoff [11]. There is presented in the following a method of development of the Type B series which is believed to be of some interest, including a necessary and sufficient condition for the convergence which is basically the same as that of Schmidt [16]. A result of Steffensen [17] is extended and shown to be related to the Charlier Type B series.

2. Statement of results. Consider the function $p(r)$, defined for $r = 0, 1, 2, \cdots$, and such that

$$\sum_{r=0}^{\infty} p(r) = 1; \qquad \sum_{r=0}^{\infty} |\, p(r) \,| = A \tag{2.1}$$

where A is some finite value. Let the n-th factorial moment be defined by

$$\begin{aligned} \mu_{(0)} &= 1 \\ \mu_{(n)} &= \sum_{r=0}^{\infty} r(r-1)(r-2)\cdots(r-n+1)p(r), \qquad (n = 1, 2, \cdots) \end{aligned} \tag{2.2}$$

For arbitrary λ let

$$\begin{aligned} L_n = \mu_{(n)} - n\mu_{(n-1)}\lambda + \frac{n(n-1)}{2!}\mu_{(n-2)}\lambda^2 \\ - \frac{n(n-1)(n-2)}{3!}\mu_{(n-3)}\lambda^3 + \cdots + (-1)^n\lambda^n. \end{aligned} \tag{2.3}$$

We prove the following results:

THEOREM. *A necessary and sufficient condition that the function $p(r)$ of* (2.1) *may be expressed by the absolutely convergent series*

$$p(r) = \frac{e^{-\lambda}\lambda^r}{r!} + L_1\frac{\partial}{\partial\lambda}\frac{e^{-\lambda}\lambda^r}{r!} + \frac{L_2}{2!}\frac{\partial^2}{\partial\lambda^2}\frac{e^{-\lambda}\lambda^r}{r!} + \cdots \tag{2.4}$$

is that

$$1 + |\, \mu_{(1)} \,| + \tfrac{1}{2!}|\, \mu_{(2)} \,| + \tfrac{1}{3!}|\, \mu_{(3)} \,| + \cdots + \frac{1}{n!}|\, \mu_{(n)} \,| + \cdots \tag{2.5}$$

converges where L_n is defined as in (2.3).

3. Generating functions. For the function $p(r)$ of (2.1) consider the generating function defined by

$$\varphi(z) = \sum_{r=0}^{\infty} z^r p(r) \tag{3.1}$$

where z is a complex variable. Because of (2.1) it is clear that the right member of (3.1) is uniformly and absolutely convergent for $|z| \leqq 1$ so that the radius of convergence of (3.1) is some value $R_1 > 1$.

The Taylor expansion of $\varphi(z)$ about the point $z = 1$ is given by

$$\varphi(z) = \varphi(1) + (z-1)\varphi'(1) + \frac{(z-1)^2}{2!}\varphi''(1) + \cdots \tag{3.2}$$

where, as may be readily obtained from (3.1),

$$\varphi^{(n)}(1) = \sum_{r=0}^{\infty} r(r-1)(r-2)\cdots(r-n+1)p(r) = \mu_{(n)}\,. \tag{3.3}$$

If it is assumed that (2.5) converges, then

$$\varphi(z) = 1 + (z-1)\mu_{(1)} + \frac{(z-1)^2}{2!}\mu_{(2)} + \cdots + \frac{(z-1)^n}{n!}\mu_{(n)} + \cdots \tag{3.4}$$

is uniformly and absolutely convergent for $|z-1| \leqq 1$.

4. Sufficiency. For arbitrary λ let us set

$$e^{-\lambda(z-1)}\left(1 + \mu_{(1)}(z-1) + \mu_{(2)}\frac{(z-1)^2}{2!} + \cdots\right) = 1 + L_1(z-1) + \frac{L_2}{2!}(z-1)^2 + \cdots \tag{4.1}$$

where the right member, because of (3.4) is absolutely convergent for $|z-1| \leqq 1$. The coefficients on the right side of (4.1) are given by

$$L_n = \mu_{(n)} - n\mu_{(n-1)}\lambda + \frac{n(n-1)}{2!}\mu_{(n-2)}\lambda^2 - \cdots + (-1)^n\lambda^n \tag{4.2}$$

and the factorial moments may also be expressed by

$$\mu_{(n)} = L_n + nL_{n-1}\lambda + \frac{n(n-1)}{2!}L_{n-2}\lambda^2 + \cdots + \lambda^n. \tag{4.3}$$

These relations are readily derived by expressing (4.1) symbolically as

$$e^{-\lambda(z-1)+\mu(z-1)} = e^{L(z-1)} \tag{4.4}$$

where after expansion μ^n and L^n are to be replaced by $\mu_{(n)}$ and L_n respectively. (Cf. Jordan [7], p. 39). From (4.1) and (3.4) there is now derived

$$\varphi(z) = e^{\lambda(z-1)}\left(1 + L_1(z-1) + \frac{L_2}{2!}(z-1)^2 + \cdots\right). \tag{4.5}$$

Since the right member of (4.5) is absolutely and uniformly convergent for $|z-1| \leqq 1$ for arbitrary λ, it may be expressed as

$$\varphi(z) = \left(1 + L_1 \frac{\partial}{\partial \lambda} + \frac{L_2}{2!}\frac{\partial^2}{\partial \lambda^2} + \cdots\right) e^{\lambda(z-1)}. \tag{4.6}$$

Since the radius of convergence of the right member of (4.6) is some value R_2 such that $|z-1| < R_2 > 1$, it may be expressed as a power series about $z = 0$, or

$$\varphi(z) = \left(1 + L_1 \frac{\partial}{\partial \lambda} + \frac{L_2}{2!}\frac{\partial^2}{\partial \lambda^2} + \cdots\right) e^{-\lambda}\left(1 + \lambda z + \frac{\lambda^2 z^2}{2!} + \cdots\right). \tag{4.7}$$

Recalling now the definition of $\varphi(z)$ as given in (3.1), there is obtained by equating coefficients of like powers of z in (3.1) and (4.7)

$$p(r) = \left(1 + L_1 \frac{\partial}{\partial \lambda} + \frac{L_2}{2!}\frac{\partial^2}{\partial \lambda^2} + \cdots\right) \frac{e^{-\lambda}\lambda^r}{r!}. \tag{4.8}$$

Since it may be readily shown that

$$\frac{\partial^n}{\partial \lambda^n}\frac{e^{-\lambda}\lambda^r}{r!} = (-1)^n \Delta^n \frac{e^{-\lambda}\lambda^r}{r!} \tag{4.9}$$

where

$$\Delta \frac{e^{-\lambda}\lambda^r}{r!} = \frac{e^{-\lambda}\lambda^r}{r!} - \frac{e^{-\lambda}\lambda^{r-1}}{(r-1)!}$$

and

$$\Delta^n \frac{e^{-\lambda}\lambda^r}{r!} = \Delta^{n-1}\frac{e^{-\lambda}\lambda^r}{r!} - \Delta^{n-1}\frac{e^{-\lambda}\lambda^{r-1}}{(r-1)!}$$

we may also write (4.8) as

$$p(r) = \frac{e^{-\lambda}\lambda^r}{r!} - L_1 \Delta \frac{e^{-\lambda}\lambda^r}{r!} + \frac{L_2}{2!}\Delta^2 \frac{e^{-\lambda}\lambda^r}{r!} - \frac{L_3}{3!}\Delta^3 \frac{e^{-\lambda}\lambda^r}{r!} + \cdots. \tag{4.10}$$

5. Necessity. Assume that the function $p(r)$ of (2.1), for arbitrary λ, is given by the absolutely convergent series

$$p(r) = \left(1 + L_1 \frac{\partial}{\partial \lambda} + \frac{L_2}{2!}\frac{\partial^2}{\partial \lambda^2} + \cdots\right) \frac{e^{-\lambda}\lambda^r}{r!}. \tag{5.1}$$

Since $e^{-\lambda}\lambda^r/r!$ is continuous with respect to λ, there follows, where z is a complex variable and $|z| \leqq 1$

$$\sum_{r=0}^{\infty} z^r p(r) = \sum_{r=0}^{\infty} \frac{z^r e^{-\lambda}\lambda^r}{r!} + L_1 \frac{\partial}{\partial \lambda}\sum_{r=0}^{\infty}\frac{z^r e^{-\lambda}\lambda^r}{r!} + \frac{L_2}{2!}\frac{\partial^2}{\partial \lambda^2}\sum_{r=0}^{\infty}\frac{z^r e^{-\lambda}\lambda^r}{r!} + \cdots$$

$$= e^{\lambda(z-1)}\left(1 + L_1(z-1) + \frac{L_2}{2!}(z-1)^2 + \cdots\right) \tag{5.2}$$

$$= 1 + M_1(z-1) + \frac{M_2}{2!}(z-1)^2 + \frac{M_3}{3!}(z-1)^3 + \cdots$$

where

$$M_n = L_n + nL_{n-1}\lambda + \frac{n(n-1)}{2!}L_{n-2}\lambda^2 + \cdots + \lambda^n. \tag{5.3}$$

From (5.2) it follows that

$$M_n = \mu_{(n)} \tag{5.4}$$

where $\mu_{(n)}$ is as defined in (3.3). Since (5.1) becomes for $r = 0, \lambda = 0$

$$1 - \mu_{(1)} + \frac{1}{2!}\mu_{(2)} - \frac{1}{3!}\mu_{(3)} + \cdots \tag{5.5}$$

the assumed absolute convergence implies that

$$1 + |\mu_{(1)}| + \frac{1}{2!}|\mu_{(2)}| + \frac{1}{3!}|\mu_{(3)}| + \cdots + \frac{1}{n!}|\mu_{(n)}| + \cdots \tag{5.6}$$

converges.

6. Remarks. Obrechkoff [11] shows that his result includes those of Pollaczek-Geiringer [12], Szegö [12] (p. 110) and Jacob [5]. His theorem states that if the function $p(r)$, $(r = 0, 1, 2, \cdots)$, satisfies the following conditions

$$\sum_{r=1}^{\infty} 2^r r^A \, | \, p(r) \, | \tag{6.1}$$

is convergent for each finite number A, and

$$\frac{(4\lambda)^n}{(n+1)!}\sum_{r=1}^{n}\frac{|\,p(r)\,|}{r}(e^{-\lambda}\lambda^r/r!)^{-1} \tag{6.2}$$

tends toward zero as n increases indefinitely then $p(r)$ may be expressed in a convergent Charlier Type B series.

Uspensky [18] shows that if

$$\sum_{r=0}^{\infty} z^r p(r) \tag{6.3}$$

has a radius of convergence $R > 2$ then $p(r)$ may be expressed in a convergent Charlier Type B series.

Schmidt [16] shows that a necessary and sufficient condition for the convergence is that the function $\varphi(z)$ defined as in (3.1) (he does not explicitly impose the condition (2.1) on $p(r)$) be regular inside the two circles $|z| < 1$ and $|z - 1| < 1$ and with all its derivatives is continuous on the peripheries also. In the case that $p(r) \geqq 0$, the condition (2.5) is stronger, in fact in this case Schmidt [16] shows that a necessary and sufficient condition is that

$$\lim_{r=\infty} p(r)2^r r^k = 0$$

for all integral $k \geqq 0$. If $p(r) \geqq 0$, then Uspensky's condition is only just enough stronger than Schmidt's to keep it from being sufficient.

If (6.1) is satisfied, or if (6.3) is satisfied then (3.1) is absolutely convergent for $|z| \leqq 2$. Therefore, the point $z = 2$ is contained in the circle of convergence of (3.2) or (3.4) which implies that

$$1 + |\mu_{(1)}| + \frac{1}{2!}|\mu_{(2)}| + \cdots + \frac{1}{n!}|\mu_{(n)}| + \cdots$$

converges.

It is deemed worthy of special mention to point out, as both Schmidt and Uspensky have done, the striking fact that the necessary and sufficient condition for the validity of (2.4) is independent of λ. This arbitrariness of λ enables us to dispose of it so as to obtain better convergence. Indeed if we set $\lambda = \mu_{(1)}$ then as is evident from (4.2) $L_1 = 0$.

7. Special cases. It is of interest to note that (4.8) is the Taylor expansion if $p(r) = e^{-\mu}\mu^r/r!$, (r = 0, 1, 2, $\cdots$), for then (4.2) becomes

$$L_n = (\mu - \lambda)^n \tag{7.1}$$

since for the Poisson Exponential Distribution $e^{-\mu}\mu^r/r!$, $(r = 0, 1, 2, \cdots)$, $\mu_{(n)} = \mu^n$ and (4.8) is then

$$\frac{e^{-\mu}\mu^r}{r!} = \frac{e^{-\lambda}\lambda^r}{r!} + (\mu - \lambda)\frac{\partial}{\partial\lambda}\frac{e^{-\lambda}\lambda^r}{r!} + \frac{(\mu - \lambda)^2}{2!}\frac{\partial^2}{\partial\lambda^2}\frac{e^{-\lambda}\lambda^r}{r!} + \cdots . \tag{7.2}$$

If $p(r)$ is finite, that is if $p(r) = 0$ for $r \geq n + 1$ then $\mu_{(k)} = 0$ for $k \geq n + 1$. Thus, for a finite function the condition (2.5) is satisfied.

8. Factorial moments. For functions $p(r)$, $(r = 0, 1, 2, \cdots)$, satisfying (2.5), there may be derived from (3.1) and (3.4) the relation

$$r!p(r) = \mu_{(r)} - \mu_{(r+1)} + \frac{1}{2!}\mu_{(r+2)} - \frac{1}{3!}\mu_{(r+3)} + \cdots, \quad (r = 0, 1, 2, \cdots), \tag{8.1}$$

since each side is $\varphi^{(r)}(0)$ derived respectively from (3.1) and (3.4). It should be noted that for $\lambda = 0$ (4.5) leads to (8.1) rather than (4.8) so that (8.1) may be considered as the Charlier Type B series for $\lambda = 0$. The result (8.1) was derived for finite functions by Steffensen [17]. (Also compare Kaplansky [8]). This may also be expressed symbolically by

$$p(r) = \mu^r e^{-\mu}/r!, \qquad (r = 0, 1, 2, \cdots), \tag{8.2}$$

where after expansion μ^n is to be replaced by $\mu_{(n)}$. It is of interest to note the relation between the symbolic expression for $p(r)$ as a Poisson Exponential in (8.2) and the series (4.8), for (4.8) may be expressed symbolically as

$$\begin{aligned} p(r) &= e^{L(\partial/\partial\lambda)} \cdot \frac{e^{-\lambda}\lambda^r}{r!} = e^{-(\lambda+L)}(\lambda + L)^r/r! \\ &= \mu^r e^{-\mu}/r! \end{aligned} \tag{8.3}$$

since $e^{a(d/dx)}f(x) = f(x + a)$ and the relations (4.2), (4.3), (4.4).

9. Illustrations. Consider the function

$$p(r) = 1/2^{r+1}, \qquad (r = 0, 1, 2, \cdots). \tag{9.1}$$

For this function

$$\varphi(z) = \sum_{r=0}^{\infty} z^r p(r) = 1/(2 - z) \tag{9.2}$$

and

$$\varphi^{(n)}(1) = \mu_{(n)} = n! \tag{9.3}$$

so that (2.5) becomes

$$1 + 1 + 1 + \cdots \tag{9.4}$$

which does not converge. (It may be of interest to note that for this case (8.1) yields

$$p(0) = 1 - 1 + 1 - 1 + 1 - \cdots . \tag{9.5}$$

The series on the right in (9.5) is not convergent but is summable C_1 to $\frac{1}{2}$. For the latter see for example R. P. Agnew, [19].) In this case the first several coefficients of (4.8) are for $\lambda = 1$,

$$\begin{aligned} &L_1 = 0, \quad \frac{L_2}{2!} = .5000, \quad \frac{L_3}{3!} = .3333, \quad \frac{L_4}{4!} = .3750 \\ &\frac{L_5}{5!} = .3667, \quad \frac{L_6}{6!} = .3681, \quad \frac{L_7}{7!} = .3679, \quad \cdots . \end{aligned} \tag{9.6}$$

Let us now consider the function

$$p(0) = \tfrac{1}{2}, \quad p(r) = \tfrac{1}{3}^r, \qquad (r = 1, 2, \cdots). \tag{9.7}$$

For this function

$$\varphi(z) = \sum_{r=0}^{\infty} z^r p(r) = \tfrac{1}{2} + \frac{z}{3-z} \tag{9.8}$$

and

$$\varphi^{(n)}(1) = \mu_{(n)} = \frac{n!}{2^n}\left(\frac{3}{2}\right), \qquad (n = 1, 2, \cdots), \tag{9.9}$$

so that (2.5) becomes

$$1 + \left(\frac{3}{2}\right)\frac{1}{2} + \left(\frac{3}{2}\right)\frac{1}{2^2} + \left(\frac{3}{2}\right)\frac{1}{2^3} + \cdots \tag{9.10}$$

which converges. For this case (8.1) yields

$$
\begin{aligned}
p(0) &= 1 - \left(\frac{3}{2}\right)\frac{1}{2} + \left(\frac{3}{2}\right)\frac{1}{2^2} - \left(\frac{3}{2}\right)\frac{1}{2^3} + \cdots = \tfrac{1}{2} \\
p(1) &= \left(\frac{3}{2}\right)\frac{1}{2} - 2!\left(\frac{3}{2}\right)\frac{1}{2^2} + \frac{3!}{2!}\left(\frac{3}{2}\right)\frac{1}{2^3} - \cdots = \tfrac{1}{3}
\end{aligned} \tag{9.11}
$$

etc.

In this case, the first several coefficients of (4.8) are for $\lambda = 0.75$

$$
L_1 = 0, \qquad \frac{L_2}{2!} = .093750, \qquad \frac{L_3}{3!} = .046875, \qquad \frac{L_4}{4!} = .019043
$$
$$
\frac{L_5}{5!} = .010840, \qquad \frac{L_6}{6!} = .005173, \qquad \frac{L_7}{7!} = .002622, \qquad \cdots . \tag{9.12}
$$

Let us now consider the function (suggested by Prof. C. Wexler)

$$
p(0) = \frac{5}{3}, \qquad p(r) = (-1)^r \frac{5}{3}\left(\frac{2}{3}\right)^r, \qquad (r = 1, 2, \cdots). \tag{9.13}
$$

For this function

$$
\sum_{r=0}^{\infty} p(r) = 1, \qquad \sum_{r=0}^{\infty} |\, p(r) \,| = 5 \tag{9.14}
$$

$$
\varphi(z) = \sum_{r=0}^{\infty} z^r\, p(r) = 5/(3 + 2z) \tag{9.15}
$$

$$
\varphi^{(n)}(1) = \mu_{(n)} = (-1)^n\, n!\, (2/5)^n. \tag{9.16}
$$

In this case (2.5) becomes

$$
1 + \frac{2}{5} + \left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^3 + \cdots \tag{9.17}
$$

which converges and (8.1) yields

$$
\begin{aligned}
p(0) &= 1 + \frac{2}{5} + \left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^3 + \cdots = 5/3 \\
p(1) &= -2/5 - 2!(2/5)^2 - \frac{3!}{2!}(2/5)^3 - \cdots = -\frac{5}{3} \cdot \frac{2}{3}
\end{aligned} \tag{9.18}
$$

etc.

Note that for this case (6.1) or (6.3) are *not* satisfied. Using $\lambda = 1$, it is found that

$$
L_1 = -1.4, \qquad \frac{L_2}{2!} = 1.06, \qquad \frac{L_3}{3!} = -.5906, \qquad \frac{L_4}{4!} = .2779, \quad \cdots . \tag{9.19}
$$

REFERENCES

[1] A. C. AITKEN, *Statistical Mathematics*, London, 1939, pp. 58–59, pp. 66–67, pp. 76–79.

[2] L. A. AROIAN, "The type B Gram-Charlier series," *Annals of Math. Stat.*, Vol. 8 (1937), p. 183.

[3] C. V. L. CHARLIER, "Über die darstellung willkürlicher Funktionen," *Arkiv. f. Math. Astron. o. Fysik*, Vol. 2 (1905–6), pp. 1–33.

[4] ARNE FISHER, *The Mathematical Theory of Probabilities*, Macmillan, 1923, pp. 271–279.

[5] M. JACOB, "Über die Charlier'sche B-Reihe," *Skand. Aktuarie-tids.*, Vol. 15(1932), pp. 286–291.

[6] CHARLES JORDAN, "Sur la probabilité des épreuves répétées, le théorème de Bernoulli et son inversion," *Bull. Soc. Math. de France*, Vol. 54 (1926), pp. 101–137.

[7] CHARLES JORDAN, *Statistique Mathématique*, Paris, 1927, p. 39.

[8] I. KAPLANSKY, "The asymptotic distribution of runs of consecutive elements," *Annals of Math. Stat.*, Vol. 16 (1945), pp. 200–203.

[9] M. KENDALL, *Advanced Theory of Statistics*, Vol. 1, Griffin, 1943. pp. 154–156.

[10] R. v. MISES, *Wahrscheinlichkeitsrechnung*, Wien. 1931, pp. 265–269.

[11] N. OBRECHKOFF, "Sur la loi de Poisson, la série de Charlier et les équations linéaires aux différences finies du premier ordre a coefficients constants," *Actualités Sci. et Ind.*, Vol. 740 (1938).

[12] H. POLLACZEK-GEIRINGER, "Die Charlier'sche Entwicklung willkürlicher Verteilungen," *Skand. Aktuarie-tids.*, Vol. 11 (1928), pp. 98–111.

[13] H. POLLACZEK-GEIRINGER, "Über die Poissonsche Verteilung und die Entwicklung willkürlicher Verteilungen," *Zeits. f. Angew. Math.*, Vol. 8 (1928), pp. 292–309.

[14] H. L. RIETZ, *Mathematical Statistics*, Chicago, 1927, pp. 60–68, pp. 170–172.

[15] E. SCHMIDT, *Sitzungsberichte der Preuss. Akad. d. Wiss.*, 1928, p. 148.

[16] E. SCHMIDT, "Über die Charlier-Jordansche Entwicklung einer willkürlichen Funktion nach der Poissonschen Funktion und ihren Ableitungen," *Zeits. Angew. Math. Mech.*, Vol. 13 (1933), pp. 139–142.

[17] J. F. STEFFENSEN, "Factorial moments and discontinuous frequency functions," *Skand. Aktuarie-tids.*, Vol. 6 (1923), pp. 73–89.

[18] J. V. USPENSKY, "On Charles Jordan's series for probability," *Annals of Math.*, Vol. 32 (1931), pp. 306–312.

[19] R. P. AGNEW, "Summability of power series," *Amer. Math. Monthly*, Vol. 53 (1946), pp. 251–259.

NOTES

This section is devoted to brief research expository articles on methodology and other short items.

ON SMALL-SAMPLE ESTIMATION

By George W. Brown

Iowa State College

1. Summary. This paper discusses some of the concepts underlying small sample estimation and reexamines, in particular, the current notions on "unbiased" estimation. Alternatives to the usual unbiased property are examined with respect to invariance under simultaneous one-to-one transformation of parameter and estimate; one of these alternatives, closely related to the maximum likelihood method, seems to be new. The property of being unbiased in the likelihood sense is essentially equivalent to the statement that the estimate is a maximum likelihood estimate based on some distribution derived by integration from the original sampling distribution, by virtue of a "hereditary" property of maximum likelihood estimation.

An exposition of maximum likelihood estimation is given in terms of optimum pairwise selection with equal weights, providing a type of rationale for small sample estimation by maximum likelihood.

2. Introduction. In large sample theory of estimation the problems are generally formulated in terms of a random variable $x = (x_1, x_2, \cdots, x_n)$ and a product distribution with, say, a density $g(x|\theta) = f(x_1|\theta)f(x_2|\theta) \cdots f(x_n|\theta)$ where n is permitted to increase without limit. For small sample theory it is sufficient to consider an arbitrary distribution, not necessarily of product form, depending on a parameter θ. For convenience we will assume a distribution density of fixed form $g(x|\theta)$, where x is in Euclidean n-space and θ in Euclidean k-space, $k \leq n$. Granting at the outset that a complete rationale for estimation must be based on considerations like those of Wald [4, 1939] dealing with specified risk functions, it is still a difficult process, in practice, to specify the risk functions and solve the ensuing mathematics problems. It may still be to the point, then, to consider general properties that estimates might be required to have in order to be considered "acceptable", or perhaps even "optimum", over a class of "acceptable" estimates.

In large-sample theory the situation is fairly simple. Consistent estimates have the property that the estimate converges in probability to the true parameter value. "Best" or "optimum" estimates are defined in terms of the order of convergence, or asymptotic variance. All reasonable definitions of "optimum" become asymptotically equivalent, since they all measure essentially the rate of

convergence, so that one might ask for least variance, or least expected absolute deviation, or least expected kth power, without affecting the optimum estimate, in general. Moreover, the consistency property and the optimum properties are in general invariant under simultaneous one-to-one transformation of the parameter and its estimate, i.e., the square of an asymptotically optimum estimate of σ will be an asymptotically optimum estimate of σ^2. Finally, a general estimation method, the method of maximum likelihood, leads to optimum estimates in large samples.

In small samples, on the other hand, the search for corresponding criteria has led to the investigation of best "unbiased" estimates, and the like, where few, if any, of the definitions discussed possess an invariance property under simultaneous one-to-one transformation of the parameter and its estimate.

3. Unbiased estimation. To ensure, in small-sample estimation, that an estimate bears some relation to the parameter it is estimating, it has become the custom to require that an estimate be *unbiased*, which means that the expected value of the estimate agrees with the parameter value. This condition was suggested by the consistency property which is required in large-sample estimation. It ensures, moreover, that the average of a large number of independent estimates made on the same basis will provide a consistent estimate, in the large sample sense. While this consistency property of the average may at times be convenient in practical situations, the fact remains that the problem of estimation from a number of such observations is a different estimation problem, the "best" solution to which need not be the average of the "best" solutions of the original problem corresponding to estimation of θ from a single observation on x, where x has a density $g(x|\theta)$. More to the point, however, is the objection that an unbiased estimate of a parameter does not in general transform into an unbiased estimate when both estimate and parameter are subjected to the same one-to-one transformation. Moreover, one can easily construct situations for which the only acceptable unbiased estimates are clearly inferior from almost any point of view, to estimates which are biased (Girshick, Mosteller and Savage, [1, 1946], and Halmos [2, 1946]).

It may be of interest to consider a few reasonable alternatives to the lack of bias requirement, which seem to accomplish as much as the conventional definition and which, in addition, have an invariance under one-to-one transformation of the parameter and estimate. To avoid confusion, let us attach the qualifying prefix "mean" to the usual unbiased property, so that an estimate will be said to be *mean-unbiased* if its expected value agrees with the parameter value.

Consider as one alternative the following property. An estimate of a one-dimensional parameter θ will be said to be *median-unbiased*, if for fixed θ, the median of the distribution of the estimate is at the value θ, i.e., the estimate underestimates just as often as it overestimates. This requirement seems for most purposes to accomplish as much as the mean-unbiased requirement and has the additional property that it is invariant under one-to-one transformation.

A different alternative requirement which is invariant under transformations is suggested by the definition of unbiased tests of significance (Neyman and Pearson [3, 1936]). Let us say that an estimate is *likelihood-unbiased* if $h(\theta \mid \theta') \leq h(\theta \mid \theta)$, where the estimate $\hat{\theta}$ has probability density $h(\hat{\theta} \mid \theta)$. In other words, an estimation method is likelihood-unbiased if estimates in the neighborhood of a given parameter value θ would occur more frequently when the true value is itself θ than when it differs from θ. On intuitive grounds this seems to be an acceptable kind of requirement, applicable to a very general class of estimation problems. It is evident that the assumption of a density plays no important role here; the situation is analogous to the maximum likelihood situation. The property itself is invariant under simultaneous one-to-one transformations of parameter and estimate for the same reason that maximum likelihood estimates are invariant under such transformations, in fact one can readily see that the likelihood-unbiased condition is equivalent to requiring that $\hat{\theta}$ have such a distribution, as a function of θ, that the maximum likelihood estimate of θ based on $\hat{\theta}$ will be actually equal to $\hat{\theta}$. The obvious implication of this fact is that if a function $\phi(x)$ is given (possibly a sufficient statistic for θ) then there is an essentially unique likelihood-unbiased estimate $\hat{\theta}$ based on ϕ, obtained by finding the maximum likelihood estimate of θ in the distribution of ϕ as a function of θ.

As an example, consider the estimation of σ^2 from a sample of n observations from a normal distribution. Let S^2 be the usual sum of squares, where S^2/σ^2 is distributed like χ^2 on $n - 1$ degrees of freedom. Then the only likelihood—unbiased estimate of σ^2 based on S^2 is $S^2/(n - 1)$. In this case $S^2/(n - 1)$ is also mean-unbiased, a fact which is normally quoted as justification for the division by $n - 1$. Curiously enough, it is customary to estimate σ by $\sqrt{S^2/(n - 1)}$, even though this is a biased estimate of σ, according to the usual notion of "unbiased", referred to here as "mean-unbiased". On the other hand, $\sqrt{S^2/(n - 1)}$ is a perfectly good likelihood-unbiased estimate of σ, by virtue of the invariance under transformations. It might be pointed out, in passing, that the estimate $S^2/(n - 1)$ does not have minimum mean square about σ^2, but that the optimum divisor for minimizing the mean square error about σ^2 is $n + 1$.

The fact that a likelihood-unbiased estimate is the maximum likelihood estimate based on the distribution of the estimate itself suggest further examination of maximum likelihood estimates. If we define a *simple* estimate as one which completely determines a probability distribution for x, then we have as a theorem, the following:

A simple maximum likelihood estimate $\hat{\theta}(x)$ is likelihood-unbiased. What this means is essentially that maximum-likelihood is "hereditary", i.e. if $\hat{\theta}(x)$ maximizes $g(x \mid \theta)$ in a space of n dimensions, and $\hat{\theta}$ has a derived density $h(\hat{\theta} \mid \theta)$ in a space of $k \leq n$ dimensions, then $\theta = \hat{\theta}$ maximizes $h(\hat{\theta} \mid \theta)$. The proof follows readily from the fact that $h(\hat{\theta} \mid \theta)$ is obtained by integration of $g(x \mid \theta)$ over all x such that $\hat{\theta}(x) = \hat{\theta}$.

The example of estimating σ^2, quoted above, shows that the word "simple" cannot be omitted from the statement above. For example, the simple estimate in the parent distribution is the joint estimate $(x, S^2/n)$ of (m, σ^2) and in fact the joint estimate is likelihood-unbiased. On the other hand, S^2/n is not a simple maximum likelihood estimate, and we observe that S^2/n is not likelihood-unbiased. $S^2/(n-1)$ is a simple maximum likelihood estimate of σ^2 based on the distribution of S^2 itself, so that $S^2/(n-1)$ is, as a result, likelihood unbiased.

One can exhibit situations in which the conventional mean-unbiased property is very unnatural, while the likelihood-unbiased property may be quite natural. Consider, for example, the case where σ^2 is to be estimated by use of a χ^2-distributed S^2 with $n-1$ degrees of freedom, but subject to the condition $\sigma^2 \geq \sigma_0^2$, where σ_0^2 is known in advance. Then the estimate $\sigma^2 = \max [S^2/(n-1), \sigma_0^2]$ is certainly biased according to conventional definitions, but is nevertheless, likelihood unbiased. To get a mean-unbiased estimate when σ^2 is near to σ_0^2 is impossible except by admitting estimates less than σ_0^2, which is clearly foolish if it is known that $\sigma^2 \geq \sigma_0^2$.

It may be of interest to include a brief discussion of maximum likelihood estimation in terms of pairwise selection of alternatives, providing a sort of optimum property for maximum likelihood estimation in small samples, in addition to the likelihood-unbiased property. Consider a choice to be made between only two alternative values of θ, say θ_0 and θ_1, by dividing the sample space into two regions S_0 and S_1, such that θ_0 is accepted when x falls in S_0 and θ_1 is accepted when x falls in S_1. Then

$$P_{\theta_0}(S_0) + P_{\theta_0}(S_1) = P_{\theta_1}(S_0) + P_{\theta_1}(S_1) = 1.$$

$P_{\theta_1}(S_0)$ is the probability of making the error of accepting θ_0 when $\theta = \theta_1$ and $1 - P_{\theta_0}(S_0)$ is the probability of making the error of accepting θ_1 when $\theta = \theta_0$. If the two errors are weighted equally, it is evident that a "best" test will choose S_0 so as to minimize $P_{\theta_1}(S_0) + 1 - P_{\theta_0}(S_0)$. It is well known that S_0 will minimize the indicated quantity if S_0 consists of all points x such that $g(x \mid \theta_0) \geq g(x \mid \theta_1)$. Thus we may speak of the region S_0 defined by $g(x \mid \theta_0) \geq g(x \mid \theta_1)$ as an *optimum equal risk acceptance region* for θ_0 against θ_1. Now if we transfer our attention to the general estimation problem we see that the maximum likelihood estimate $\hat{\theta}(x)$ is that value of θ which would be accepted by the optimum equal risk acceptance procedure against all other θ's.

REFERENCES

[1] M. A. Girshick, Frederick Mosteller, and L. J. Savage, "Unbiased estimates for certain binomial sampling problems with applications," *Annals of Math. Stat.*, Vol. 17 (1946), p. 13.

[2] Paul R. Halmos, "The theory of unbiased estimation," *Annals of Math. Stat.*, Vol. 17 (1946), p. 34.

[3] J. Neyman and E. S. Pearson, "Unbiased critical regions of Type A and Type A_1", *Stat. Res. Mem.*, Vol. 1, p. 1.

[4] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses", *Annals of Math. Stat.*, Vol. 10 (1939), p. 299.

A NOTE ON REGRESSION ANALYSIS

By Abraham Wald
Columbia University

1. Introduction. In regression analysis a set of variables $y, x_1, \cdots, x_p$ is considered where y is called the dependent variable and $x_1, \cdots, x_p$ are the independent variables. Let y_α denote the αth observation on y and $x_{i\alpha}$ the αth observation on x_i, $(i = 1, \cdots, p; \alpha = 1, \cdots, N)$. The observations $x_{i\alpha}$ are treated as given constants, while the observations $y_1, \cdots, y_N$ are regarded as chance variables. The following two assumptions are usually made concerning the joint distribution of the variates $y_1, \cdots, y_N$:

(a) The variates $y_1, \cdots, y_N$ are normally and independently distributed with a common unknown variance σ^2.

(b) The expected value of y_α is equal to $\beta_1 x_{1\alpha} + \cdots + \beta_p x_{p\alpha}$ where $\beta_1, \cdots, \beta_p$ are unknown constants.

In some problems it seems reasonable to assume that the regression coefficients $\beta_1, \cdots, \beta_p$ are not constants, but chance variables. This leads to a different probability model for regression analysis and the object of this note is to discuss certain aspects of this model. In what follows in this note we shall make the following assumptions concerning the joint distribution of the chance variables $y_1, \cdots, y_N; \beta_1, \cdots, \beta_p$.

Assumption 1. For given values of $\beta_1, \cdots, \beta_p$ the joint conditional probability density function of $y_1, \cdots, y_N$ is given by

$$\frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\sum_{\alpha=1}^{N}(y_\alpha - \beta_1 x_{1\alpha} - \cdots - \beta_p x_{p\alpha})^2\right] \tag{1.1}$$

Assumption 2. The regression coefficients $\beta_1, \cdots, \beta_p$ are independently distributed.

Assumption 3. The regression coefficients $\beta_1, \cdots, \beta_r$, $(r \leq p)$, are normally distributed with zero means and a common variance σ'^2.

The purpose of this note is to derive confidence limits for the ratio $\frac{\sigma'^2}{\sigma^2}$. Such confidence limits have been derived by the author [1] for analysis of variance problems assuming that there are only main effects but no interactions. The regression problem treated in the present note is much more general and includes all the analysis of variance problems with or without interactions as special cases.

It should be remarked that Assumptions 2 and 3 do not exclude the case where $\beta_{r+1}, \cdots, \beta_p$ are constants.

2. Derivation of confidence limits for the ratio $\frac{\sigma'^2}{\sigma^2}$. Let $b_1, \cdots, b_p$ be the sample estimates of $\beta_1, \cdots, \beta_p$ obtained by the method of least squares. We

shall denote the difference $b_i - \beta_i$ by ϵ_i, $(i = 1, \cdots, p)$. It is known that for given values of $\beta_1, \cdots, \beta_p$ the conditional joint distribution of $\epsilon_1, \cdots, \epsilon_p$ is normal with zero means and variance-covariance matrix $\| c_{ij} \| \sigma^2$ where

$$\| c_{ij} \| = \| a_{ij} \|^{-1} \tag{2.1}$$

and

$$a_{ij} = \sum_{\alpha=1}^{N} x_{i\alpha} x_{j\alpha}, \qquad (i, j = 1, \cdots, p). \tag{2.2}$$

Since the conditional distribution of $\epsilon_1, \cdots, \epsilon_p$ does not depend on the values of $\beta_1, \cdots, \beta_p$, the unconditional distribution of $\epsilon_1, \cdots, \epsilon_p$ is the same as the conditional one, and the set of variates $(\beta_1, \cdots, \beta_p)$ is independently distributed of the set $(\epsilon_1, \cdots, \epsilon_p)$. From this and Assumptions 2 and 3 it follows that $b_1, \cdots, b_r$ have a joint normal distribution and that

$$Eb_i = 0, \qquad (i = 1, \cdots, r) \tag{2.3}$$

and

$$Eb_i b_j = \left(c_{ij} + \delta_{ij} \frac{\sigma'^2}{\sigma^2} \right) \sigma^2, \qquad (i, j = 1, \cdots, r) \tag{2.4}$$

where $\delta_{ij} = 0$ for $i \neq j$ and $= 1$ for $i = j$.

We shall denote $\frac{\sigma'^2}{\sigma^2}$ by λ and the elements of the inverse of $\| c_{ij} + \delta_{ij}\lambda \|$ by $d_{ij}(\lambda)$, i.e.,

$$\| d_{ij}(\lambda) \| = \| c_{ij} + \delta_{ij}\lambda \|^{-1}, \qquad (i, j = 1, \cdots, r). \tag{2.5}$$

Then the quadratic form

$$Q(\lambda) = \frac{1}{\sigma^2} \sum_{j=1}^{r} \sum_{i=1}^{r} d_{ij}(\lambda) b_i b_j \tag{2.6}$$

has the χ^2 distribution with r degrees of freedom.

It is known that for any given values of $\beta_1, \cdots, \beta_p, b_1, \cdots, b_p$ the quadratic form

$$Q_a = \frac{1}{\sigma^2} \sum_{\alpha=1}^{N} (y_\alpha - b_1 x_{1\alpha} - \cdots - b_p x_{p\alpha})^2 \tag{2.7}$$

has the χ^2 distribution with $N - p$ degrees of freedom provided that the rank of the matrix $\| x_{i\alpha} \|$ is p. Hence Q_a and $Q(\lambda)$ are independently distributed and the ratio

$$F = \frac{N - p}{r} \frac{Q(\lambda)}{Q_a} \tag{2.8}$$

has the F-distribution with r and $N - p$ degrees of freedom.

Let F_1 and F_2 be two values chosen so that

$$\text{Prob.} \{F_1 \leqq F \leqq F_2\} = c \tag{2.9}$$

where c is a given positive constant less than 1. Then the set of all values λ for which the inequality

$$F_1 \leq \frac{N-p}{r}\frac{Q(\lambda)}{Q_a} \leq F_2 \tag{2.10}$$

holds forms a confidence set for λ with the confidence coefficient c.

We shall now show that $Q(\lambda)$ is a monotonic function of λ and, therefore, the confidence set determined by (2.10) is an interval. Let $\| g_{ij} \|$, $(i, j = 1, \cdots, r)$, be an orthogonal matrix and let

$$b_i^* = \sum_{j=1}^{r} g_{ij} b_j . \tag{2.11}$$

It then follows from (2.3) and (2.4) that

$$E(b_i^*) = 0, \qquad (i = 1, \cdots, r) \tag{2.12}$$

and

$$E(b_i^* b_j^*) = (c_{ij}^* + \delta_{ij}\lambda)\sigma^2, \qquad (i, j = 1, \cdots, r) \tag{2.13}$$

where

$$c_{ij}^* = \sum_{l=1}^{r}\sum_{k=1}^{r} g_{ik} g_{jl} c_{kl}' . \tag{2.14}$$

Let

$$\| d_{ij}^*(\lambda) \| = \| c_{ij}^* + \delta_{ij}\lambda \|^{-1}, \qquad (i, j = 1, \cdots, r) \tag{2.15}$$

and put

$$Q^*(\lambda) = \frac{1}{\sigma^2} \Sigma\Sigma\, d_{ij}^*(\lambda) b_i^* b_j^* .$$

It is easy to verify that $Q^*(\lambda)$ is identically equal to $Q(\lambda)$. Hence, to prove the monotonicity of $Q(\lambda)$, it is sufficient to show that $Q^*(\lambda)$ is a monotonic function of λ. Since no restrictions as to the choice of the orthogonal matrix $\| g_{ij} \|$ are made, we shall choose it so that the matrix $\| c_{ij}^* \|$ becomes diagonal, i.e., $c_{ij}^* = 0$ for $i \neq j$, $(i, j = 1, \cdots, r)$. Then

$$d_{ij}^*(\lambda) = 0 \qquad \text{for } i \neq j \tag{2.16}$$

and

$$d_{ii}^*(\lambda) = \frac{1}{c_{ii}^* + \lambda} . \tag{2.17}$$

Hence

$$Q(\lambda) = Q^*(\lambda) = \frac{1}{\sigma^2}\sum_{i=1}^{r} \frac{b_i^{*2}}{c_{ii}^* + \lambda} \tag{2.18}$$

is a monotonically decreasing function of λ. The confidence set determined by (2.10) is, therefore, an interval.

The upper end point of the confidence interval is the root in λ of the equation

$$\frac{N-p}{r}\frac{Q(\lambda)}{Q_a} = F_1 \tag{2.19}$$

and the lower end point is the root in λ of the equation

$$\frac{N-p}{r}\frac{Q(\lambda)}{Q_a} = F_2. \tag{2.20}$$

If equation (2.20) has no root, the lower end point of the confidence interval is put equal to zero.

REFERENCE

[1] A. WALD, "On the analysis of variance in case of multiple classifications with unequal class frequencies", *Annals. of Math. Stat.*, Vol. 12 (1941).

ON THE SHAPE OF THE ANGULAR CASE OF CAUCHY'S DISTRIBUTION CURVES

BY AUREL WINTNER

The Johns Hopkins University

1. Let ξ be a *linear* random variable, that is, a random variable capable of values x represented by points of a line $-\infty < x < \infty$, and suppose, for simplicity, that ξ has a density of probability, $f(x)$. Then, subject to provisos of convergence, the series

$$F(x) = \sum_{n=-\infty}^{\infty} f(x+n)$$

represents a periodic function, of period 1, having the following significance: $F(x)$ is the density of probability of the *angular* random variable, say Ξ, which is obtained if all the states

$$\cdots, \quad \xi - 2, \quad \xi - 1, \quad \xi, \quad \xi + 1, \quad \xi + 2, \cdots$$

of the linear random variable are identified.

In other words, if a circle of unit circumference rolls from $-\infty$ to ∞ on the ξ-line, then every point of the circumference collects the various densities of probability attached to congruent points of the ξ-line, and a state of Ξ represents a point of the circumference. For a detailed study of the mapping $\xi \to \Xi$ or $f \to F$, cf. [2].

According to Poisson's summation formula, the Fourier constants of the periodic function $F(x)$ can be obtained by restricting u in $g(u)$ to an equidistant sequence of discrete values, where $g(u)$ denotes the Fourier transform of $f(x)$; cf., e.g., [5], p. 78 or [9], pp. 477–478.

2. Consider, in particular, the case in which $f(x)$ is the density of a symmetric distribution which is stable in Cauchy's sense. The determination of the totality of these linear densities of probability is due to Lévy [6]. It was shown in [8] that every such $f(x) = f(-x)$ is a decreasing function of $|x|$. As explained in [8], p. 70, this fact makes superfluous one of the axioms occurring in Gauss' postulational approach to "errors of observation."

The purpose of the present note is the deduction of the angular analogue of the fact just quoted. The analogue states that, if $f(x)$ is symmetric and stable, then the corresponding periodic $F(x)$ is decreasing for $0 \leqq x \leqq \frac{1}{2}$ (and so, for reasons of symmetry, is increasing for $\frac{1}{2} \leqq x \leqq 1$). This is contained in the italicized statement of §4 below.

In view of Poisson's rule, quoted above, the periodic densities in question can be defined by certain Fourier series representing generalizations of elliptic theta-series. From this point of view, not even the existence (i.e., the *positivity*) of the periodic densities is obvious, if arbitrary values of the "precision constant" (denoted below by q) are allowed. The difficulties involved are explained in §3.

3. If q and λ are positive constants the first of which is less than 1, then the (even, periodic) function

$$\theta_\lambda(x;q) = 1 + 2\sum_{n=1}^{\infty} q^{n^\lambda} \cos nx, \tag{1}$$

where $q^{n^\lambda} > 0$, has derivatives of arbitrarily high order at every real x. It is regular-analytic at every real x if and only if $\lambda > 0$ is replaced by $\lambda \geqq 1$, where the sign of equality holds if and only if the analytic continuation (from the x-axis) is not an entire function. In fact, it is known that a Fourier series $\Sigma(a_n \cos nx + b_n \sin nx)$ is that of a function which is regular-analytic at every real x, and has the period 2π, if and only if $|a_n| + |b_n|$ is majorized by a constant multiple of the nth power of a positive constant which is less than 1; and that the latter constant can be chosen arbitrarily small if and only if the analytic continuation does not lead to any singularity (at a $z \neq \infty$).

Since the function (1) tends to 1 uniformly in x as $q \to +0$, if λ is fixed, there belongs to every $\lambda > 0$ a positive $q^* = q^*(\lambda)$ having the property that

$$\theta_\lambda(x;q) > 0 \text{ for } 0 \leqq x < 2\pi \tag{2}$$

if $0 < q < q^*(\lambda)$. It is less obvious that, if q is sufficiently small with reference to λ, say if $0 < q < q^{**}(\lambda)$, then

$$\theta_\lambda(x;q) \text{ is decreasing for } 0 \leqq x \leqq \pi \tag{3}$$

(hence, increasing for $\pi \leqq x < 2\pi$). The existence of such a $q^{**}(\lambda) < \infty$ for every $\lambda > 0$ can be assured as follows:

If $s_n(x)$ denotes the nth partial sum of the Fourier series $\Sigma(\sin nx)/n$, then $s_n(x)$ is positive for $0 < x < \pi$ (Gronwall, Jackson; for a short proof, cf. [4]).

Hence, a partial summation shows that the sum of a sine series, $\Sigma b_n \sin nx$, must be positive for $0 < x < \pi$ if

$$nb_n - (n+1)b_{n+1} > 0 \text{ and } nb_n \to 0.$$

Since the first derivative of (1) (with respect to x) results by choosing $b_n = -2nq^{n^\lambda}$, it follows that (3) must be true if

$$n^2 q^{n^\lambda} - (n+1)^2 q^{(n+1)^\lambda} > 0$$

holds for $n = 1, 2, \cdots$. But the last inequality is readily seen to be satisfied from $n = 1$ onward if, while λ is fixed, q tends to 0. This proves that $q^{**}(\lambda)$ exists for every $\lambda > 0$.

4. From these deductions alone, it is quite unexpected that (the best values of) both $q^*(\lambda)$ and $q^{**}(\lambda)$ turn out to be independent of λ when

$$0 < \lambda \leqq 2, \tag{4}$$

i.e., that (1) *satisfies both* (2) *and* (3) *for* $0 < q < 1$, *if* (4) *is assumed.* This fact is of statistical significance, since, on the one hand, it is precisely the restriction (4) which is necessary and sufficient for the existence of Cauchy's (symmetric) "stable" distributions (cf. [6], pp. 254–263) and, on the other hand, the reduction (mod 2π) of the densities of these *linear* distributions leads to the functions (1) as angular densities (cf. [9], pp. 477–478); the numerical value of $q(< 1)$ being determined by the "precision" or "dispersion" of the resulting *angular* distributions.

Under the necessary restriction (4), the *linear* analogue of $q^*(\lambda) = 1$ and of $q^{**}(\lambda) = 1$ was proved in [6], pp. 258–263 and in [8], pp. 71–77, respectively. It will remain undecided whether the restriction (4) is necessary in either of the *angular* cases.

5. Suppose that λ has a fixed value in the range (4). Then there exists a monotone function of t, say $\alpha_\lambda(t)$, for which

$$\exp(-u^\lambda) = \int_0^\infty \exp(-u^2 t)\, d\alpha_\lambda(t)$$

is an identity in u, where $0 < u < \infty$ (cf. [1], p. 769, where further references will be found). Hence, a change of variables shows that

$$q^{n^\lambda} = \int_0^\infty q^{tn^2}\, d\alpha_\lambda(t \mid \log q \mid^{1-2\lambda})$$

is an identity in q and n, where $0 < q < 1$ and $n = 0, 1, 2, \cdots$ (the integration variable is t). Consequently from (1),

$$\theta_\lambda(x; q) = \int_0^\infty \theta_2(x; q^t)\, d\alpha_\lambda(t \mid \log q \mid^{1-2/\lambda}),$$

where $0 < q < 1$ and $-\infty < x < \infty$. In fact, the legitimacy of the term-by-term integration is obvious from $0 < q < 1$ and $d\alpha_\lambda \geqq 0$ (even though the integrals are improper).

6. Since α_λ is a non-decreasing function, it is clear from the last formula line that both (2) and (3) will be proved for $0 < q < 1$ and for every λ (satisfying (4)), if it is ascertained that both (2) and (3) hold for $0 < q < 1$ when $\lambda = 2$. But the case $\lambda = 2$ of (1) is an elliptic theta-function, for which both properties in question (cf. the diagram in [3], p. 44) are known; a simple proof can be concluded from what, in Hecke's terminology, is the Eulerian factorization of $\theta_2(x\,;q)$, as follows:

According to Jacobi, the factorization of the case $\lambda = 2$ of (1) is

$$\theta_2(x\,;q) = \prod_{n=1}^{\infty}(1 - q^{2n})(1 + 2q^{2n-1}\cos x + q^{4n-2})$$

(cf. [7], pp. 64–65). Thus

$$\theta_2(x\,;q) = c_q \prod_{n=1}^{\infty} P(x + \pi\,;q^{2n-1}),$$

where

$$c_q = \prod_{n=1}^{\infty}(1 - q^{2n})$$

and

$$P(x\,;r) = 1 - 2r\cos x + r^2, \qquad (0 < r < 1), \tag{5}$$

hence

$$P(x\,;r) > 0 \qquad (0 < r < 1).$$

Since $0 < q < 1$, this proves the case $\lambda = 2$ of (2). Furthermore, logarithmic differentiation of the product representation of $\theta_2(x\,;q)$ gives

$$\theta_2'(x\,;q) = \theta_2(x\,;q)\sum_{n=1}^{\infty} P'(x + \pi\,;q^{2n-1})/P(x + \pi\,;q^{2n-1}),$$

where $f' = df/dx$; so that, by (5),

$$P'(x + \pi\,;r) = -2r\sin x.$$

Since $0 < q < 1$, the last three formula lines and the case $\lambda = 2$ of (2) imply that

$$\theta_2'(x\,;q) < 0 \text{ if } 0 < x < \pi,$$

as claimed by the case $\lambda = 2$ of (3).

This completes the proof of the italicized assertion.

REFERENCES

[1] P. HARTMAN AND A. WINTNER, "On the spherical approach to the normal distribution law," *Amer. Jour. of Math.*, Vol. 62 (1940), pp. 759–779.

[2] E. K. HAVILAND, "On the distribution functions of the reciprocal of a function and of a function reduced mod 1," *Amer. Jour. of Math.*, Vol. 68 (1941), pp. 408–414.

[3] E. JAHNKE AND F. EMDE, *Tables of Functions with Formulae and Curves*, 3d ed. (1938), reprinted by G. E. Stechert and Co., New York.

[4] E. LANDAU, "Ueber eine trigonometrische Ungleichung," *Math. Zeit.*, Vol. 37 (1933), p. 36.

[5] H. LEBESGUE, *Leçons sur les séries trigonométriques*, Paris, 1905.

[6] P. LÉVY, *Calcul des probabilités*, Paris, 1925.

[7] H. WEBER, *Elliptische Funktionen und algebraische Zahlen*, Braunschweig, 1891.

[8] A. WINTNER, "On a class of Fourier transforms," *Amer. Jour. of Math.*, Vol. 58 (1936), pp. 45–90 and p. 425.

[9] F. ZERNIKE, "Wahrscheinlichkeitsrechnung und mathematische Statistik," *Handbuch der Physik*, Vol. 3 (1928), pp. 419–492.

A NOTE ON THE FUNDAMENTAL IDENTITY OF SEQUENTIAL ANALYSIS

BY G. E. ALBERT

U. S. Naval Ordnance Plant, Indianapolis

1. Introduction. Let $\{z_i\}$, $(i = 1, 2, 3, \cdots)$, be a sequence of real valued random variables identically distributed according to the cumulative distribution function $F(z)$. Define the sums $Z_N = z_1 + z_2 + \cdots + z_N$ for every positive integer N. Choose two positive constants a and b and define the random variable n as the smallest integer N for which one of the inequalities $Z_N \geqq a$ or $Z_N \leqq -b$ holds. The notations $P(u \mid F)$ and $E(u \mid F)$ will denote the probability of u and its expectation respectively assuming that F is the distribution of the z_i.

Wald [1] has established the results contained in the following lemmas.

LEMMA 1. *If the variance of $F(z)$ is positive, $P(n < \infty \mid F)$ equals one.*

LEMMA 2. *If there exists a positive number δ such that $P(e^z < 1 - \delta \mid F) > 0$ and $P(e^z > 1 + \delta \mid F) > 0$ and if the moment generating function $\varphi(t) = E(e^{zt} \mid F)$ exists for all real values of t, then $\varphi(t)$ has one and only one minimum at some finite value $t = t_0$. Moreover, $\varphi''(t) > 0$ for all real values of t.*

It is the purpose of this note to establish the following extension of the validity of certain results given by Wald [1], [2].

THEOREM.[1] *Under the conditions of Lemma 2 the identity*

$$E\{e^{Z_n t}[\varphi(t)]^{-n} \mid F\} = 1 \tag{1}$$

[1] Wald's results show (1) to be valid for all complex t in the domain over which $|\varphi(t)| \geqq 1$ and the validity of the differentiation clause for all real t in that domain. The importance of the present extension arises from the fact that, if $E(x \mid F) \neq 0$, then $0 < \varphi(t) < 1$ on a certain interval of the real axis.

is valid and may be differentiated with respect to t under the expectation sign any number of times for all real values of t.

PROOF. The notation t_0 will be used consistently to denote the t value at which $\varphi(t)$ has its minimum.

The proof of the theorem follows Wald's methods quite closely and certain of the results given in [1] and [2] will be used here without discussion.

Consider first the validity of (1). For an arbitrary positive integer N let P_N be the probability $P(n \leqq N \mid F)$ and let $E_N(u \mid F)$ and $E_N^*(u \mid F)$ denote the conditional expectations of u subject to the respective conditions $n \leqq N$ and $n > N$. Wald [1] has shown that for any finite real value of t

$$P_N E_N\{e^{Z_n t}[\varphi(t)]^{-n} \mid F\} + (1 - P_N)[\varphi(t)]^{-N} E_N^*\{e^{Z_N t} \mid F\} = 1. \tag{2}$$

Since $\lim_{N=\infty} P_N E_N\{[\varphi(t)]^{-n}\exp(Z_n t)\}$ is the left member of the identity (1), it suffices to demonstrate that

$$\lim_{N=\infty} (1 - P_N)[\varphi(t)]^{-N} E_N^*\{e^{Z_N t} \mid F\} = 0 \tag{3}$$

for all real values of t.

Since $1 - P_N$ tends to zero with increasing N and the expected value E_N^* involved in (3) is bounded independently of N for any fixed t, the only source of difficulty in proving (3) lies in the fact that $\varphi(t)$ may be less than unity on an interval of the real axis. That difficulty is easily avoided by the following device. Define the function

$$G(x) = [\varphi(t_0)]^{-1}\int_{-\infty}^{x} e^{zt_0}\, dF(z). \tag{4}$$

Obviously $G(x)$ is a distribution function whose moment generating function $\psi(t)$ exists for all real t. Its mean is zero and its variance is positive as will be seen from the equations $E(x \mid G) = \varphi'(t_0)/\varphi(t_0)$ and $E(x^2 \mid G) = \varphi''(t_0)/\varphi(t_0)$. It follows that $\psi(t)$ is never less than unity for real values of t.

Let Ω denote the space of all $z_1, \cdots, z_N$ and let $\Omega(n > N)$ be that subset of Ω on which $n > N$. One has

$$(1 - P_N)[\varphi(t)]^{-N} E_N^*\{e^{Z_N t} \mid F\}$$

$$= \frac{\int_{\Omega(n>N)} e^{Z_N t}\, dF(z_1)\cdots dF(z_N)}{\int_{\Omega} e^{Z_N t}\, dF(z_1)\cdots dF(z_N)} = \frac{\int_{\Omega(n>N)} e^{Z_N(t-t_0)}\, dG(z_1)\cdots dG(z_N)}{\int_{\Omega} e^{Z_N(t-t_0)}\, dG(z_1)\cdots dG(z_N)}$$

$$= (1 - Q_N)[\psi(s)]^{-N} E_N^*\{e^{Z_N s} \mid G\}$$

where $s = t - t_0$ and $Q_N = P(n \leqq N \mid G)$. By Lemma 1, $1 - Q_N$ tends to zero as N is increased. Thus, since $\psi(s) \geqq 1$ for all real t and the expected value $E_N^*\{e^{Z_N s} \mid G\}$ is bounded independently of N for a fixed t, the equation (3) holds for all real t.

The differentiability clause of the theorem requires the following modification of a very powerful theorem due to Charles Stein [3].

LEMMA 3. *Under the conditions of Lemma 2, if the minimum $\varphi(t_0)$ of $\varphi(t)$ is less than unity, there exists a positive number t_1 such that*

$$E\{\exp [nt_1 - n \log \varphi(t_0)] \mid F\} < \infty. \tag{5}$$

PROOF. If G is the distribution of the z_i, by Stein's theorem there exists a positive number t_1 such that $E(e^{nt_1} \mid G)$ is finite. Let $\Omega(n = N)$ denote the subset of Ω on which $n = N$. Then

$$\begin{aligned} P(n = N \mid G) &= \int_{\Omega(n=N)} dG(z_1) \cdots dG(z_N) \\ &= [\varphi(t_0)]^{-N} \int_{\Omega(n=N)} e^{t_0 z_N} \, dF(z_1) \cdots dF(z_N) \\ &\geqq P(n = N \mid F) \exp [\min\{at_0, -bt_0\} - N \log \varphi(t_0)]. \end{aligned}$$

It follows that

$$E\{\exp [nt_1 - n \log \varphi(t_0)] \mid F\} \leqq E\{e^{nt_1} \mid G\} \exp [-\min\{at_0, -bt_0\}]$$

and the lemma is proved.

To continue with the theorem, Wald's proof [2] suffices for the case in which $\varphi(t_0) \geqq 1$. Attention will be given only to the case $\varphi(t_0) < 1$. As pointed out in section 2 of [2], the differentiability clause of the theorem will be established if it can be shown that for any finite interval I of the real axis and any pair of integers r_1 and r_2 there exists a function $D_{r_1 r_2}(Z_n, n)$ such that for all t in I one has

$$D_{r_1 r_2}(Z_n, n) \geqq | n^{r_1} Z_n^{r_2} e^{Z_n t} [\varphi(t)]^{-n} | \tag{6}$$

and

$$E\{D_{r_1 r_2}(Z_n, n) \mid F\} < \infty. \tag{7}$$

On referring to Wald's proof and using the inequality $-\log \varphi(t) \leqq -\log \varphi(t_0)$ for all t in I, it is seen that there exists a constant C and a positive number t_2 such that the function

$$D_{r_1 r_2}(Z_n, n) \equiv C n^{r_1} [\varphi(t_0)]^{-n} (e^{Z_n t_2} + e^{-Z_n t_2})$$

satisfies (6) for all t in I. To establish (7) use the inequalities (2.4) and (2.6) in Wald [2] to obtain

$$\begin{aligned} &E\{D_{r_1 r_2}(Z_n, n) \mid F\} \\ &= C \sum_{N=1}^{\infty} P(n = N \mid F) N^{r_1} [\varphi(t_0)]^{-N} E_{n=N} \{e^{Z_n t_2} + e^{-Z_n t_2} \mid F\} \qquad (8) \\ &\leqq C\{e^{at_2} l(t_2) + e^{-bt_2} l(-t_2)\} E\{\exp [r_1 \log n - n \log \varphi(t_0)] \mid F\}. \end{aligned}$$

That (7) is indeed satisfied now follows from (5) and the finiteness of the function $l(t)$ since for a large enough integer M one has

$$\sum_{N=M}^{\infty} P(n = N \mid F) \exp [r_1 \log N - N \log \varphi(t_0)]$$

$$\leqq \sum_{N=M}^{\infty} P(n = N \mid F) \exp [Nt_1 - N \log \varphi(t_0)] < \infty.$$

Thus the expected value on the extreme right in (8) is finite. This completes the proof of the theorem.

REFERENCES

[1] A. WALD, "On cumulative sums of random variables," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 283–285.

[2] A. WALD, "Differentiation under the expectation sign in the fundamental identity of sequential analysis," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 493–496.

[3] CHARLES STEIN, "A note on cumulative sums," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 498–499.

A SIGNIFICANCE TEST AND ESTIMATION IN THE CASE OF EXPONENTIAL REGRESSION

BY D. S. VILLARS[1]

United States Rubber Company, Passaic, N. J.

1. Introduction. The principal problem under consideration in this note may be described as follows. Consider a variate, z, whose distribution for a given value of a fixed variate, t, is:

$$f(z \mid t) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(z-a+be^{-kt})^2/2\sigma^2} \tag{1.1}$$

where a, b, and k are real-valued parameters. The regression of z on t is exponential, for it follows from (1.1) that the expected value of z, given t, is:

$$E(z \mid t) = a - be^{-kt}. \tag{1.2}$$

On the basis of a random sample $0_N(z_1, t_1; z_2, t_2; \cdots; z_N, t_N)$ it is desired to test whether $k = 0$ or ∞. The problem of "fitting" a curve, $z = a - be^{-kt}$, to the sample (*i. e.* of estimating a, b, and k from the sample) will also be treated.

As an illustration of how the statistical problems described above arise in

[1] Present address, Jersey City Junior College, Jersey City, N. J.

practice, let us consider a typical situation in industrial chemistry. Let the quantity, z, be a property of a latex and let the quantity, t, be time. Suppose, furthermore, that measurement of t is without error but that measurement of z is subject to error; let it be assumed that the observed value in a measurement of z is a variate having a normal (Gaussian) distribution about the "true value," $E(z)$. On basis of N independent measurements, $z_1, z_2, \cdots, z_N$ of z at times, $t_1, t_2, \cdots, t_N$, respectively, the experimenter may wish to test the hypothesis that $k = 0$ or ∞. If this hypothesis is true the suspected exponential relation between z and t does not hold; in this case $E(z)$ is a constant ($a - b$, or a) and estimation of the constant from the data is quite straightforward. If the data conflict with the hypothesis that $k = 0$ or ∞, the experimenter may wish to estimate the parameters, a, b, and k (*i. e.*, "fit" the curve, $z = a - be^{-kt}$, to the data).

The problems considered in this note will be treated only for the case where N is an even integer (≥ 6) and the times $t_1, t_2, \cdots, t_N$ at which measurements of z are made are such that

$$t_{2\alpha} - t_{2\alpha-1} = \Delta, \qquad \text{a constant, } (\alpha = 1, 2, \cdots, n = N/2). \tag{1.3}$$

The odd time intervals, $t_3 - t_2, t_5 - t_4$, etc. do not have to be equal.

2. Test of the hypothesis that $k = 0$ or ∞. The space, say Ω, of admissible values of the parameters in (1.1) is: $\sigma^2 > 0, -\infty < a, b, k < +\infty$. Under the null hypothesis the admissible values of the parameters lie in a subspace of Ω, say ω, specified as follows: $\sigma^2 > 0, -\infty < a, b < +\infty, k = 0$, or ∞.

Let $y_j = z_{2\alpha}$ and $x_j = z_{2\alpha-1}$, $(\alpha = 1, \cdots, n = N/2)$. From (1.1) and (1.3) it follows that the n pairs x_j, y_j are normally and independently distributed with common variance, σ^2, that x_j and y_j are independent ($j = 1, 2, \cdots, n$), and that

$$\nu_j = h + m\mu_j \tag{2.1}$$

where $\nu_j = E(y_j)$, $\mu_j = E(x_j)$, $h = a(1 - e^{-k\Delta})$, and $m = e^{-k\Delta}$. The space, Ω', of admissible values of the parameters in the joint distribution of x_j, y_j, $(j = 1, \cdots, n)$, is: $\sigma^2 > 0$, $\nu_j = h + m\mu_j$, $-\infty < h < +\infty$, $-\infty < \mu_j, \nu_j < +\infty$; $0 \leq m < \infty$. The subspace of Ω', say ω', associated with the null hypothesis is: $\sigma^2 > 0$, $\nu_j = \mu_j = c$, where $c = a - b$ or a according as $k = 0$ or ∞. In Ω', the expected values of x and y lie on a line; in ω' they lie in a single point. It is clear that by transforming the original sample $0_N(z_1, t_1, \cdots, z_N, t_N)$ to a sample $0_n(x_1, y_1; \cdots; x_n, y_n)$ we have reduced the original problem to the familiar problem of linear regression in which there is "error in both variates".

The slope of the "line of best fit" to the sample points $(x_1, y_1; \cdots; x_n, y_n)$ is [1]:

$$\hat{m} = [S_{yy} - S_{xx} + \sqrt{(S_{yy} - S_{xx})^2 + 4S_{xy}^2}]/2S_{xy} \tag{2.2}$$

where

$$S_{xx} \equiv \sum_1^n (x_j - \bar{x})^2$$

$$S_{xy} \equiv \sum_1^r (x_j - \bar{x})(y_j - \bar{y})$$

$$S_{yy} \equiv \sum_1^n (y_j - \bar{y})^2$$

$$\bar{x} \equiv \sum_1^n x_j/n$$

$$\bar{y} \equiv \sum_1^n y_j/n$$

($\hat{m}$ is an estimate of m in (2.1)). Since $m = e^{-k\Delta}$ (where k and Δ are real), it is intuitively clear that when m is non-positive the sample 0_n does not conflict with the null hypothesis. The null hypothesis can be tested by means of the statistic [2, 144]

$$F' = \frac{S_{xx} + 2mS_{xy} + m^2 S_{yy}}{m^2 S_{xx} - 2mS_{xy} + S_{yy}}. \tag{2.3}$$

The null hypothesis is rejected if $\hat{m}$ is positive and F' is large. Percentage points of the distribution of F' are given in [2, 146] for $n = 3$ (1) 15 (5) 30, 40, 60, 120 and for significance levels, 0.001, .01, .05, .10, and .20. These significance levels, however, were computed for use in cases where the sign of $\hat{m}$ was irrelevant. It happens that to test the null hypothesis under consideration in this problem at a significance level α we should use a critical value of F' (given in [2]) corresponding to a significance level 2α. The reason for this is that when the null hypothesis is true the quantities m and F' are independent and the probability that $\hat{m}$ is positive is $\frac{1}{2}$—thus the chance of rejecting the null hypothesis is $\frac{1}{2}(2\alpha) = \alpha$.

3. Estimation of a, b, and k. If the data do not support the hypothesis that $k = 0$ or ∞, the experimenter may wish to estimate a, b, and k. General alternative methods of estimating these parameters will now be considered.

(1) Estimate a, b, and k from 0_N by the method of least squares; *i.e.*, solve the simultaneous equations $\partial S/\partial a = 0$, $\partial S/\partial b = 0$, and $\partial S/\partial k = 0$ for a, b, and k, where

$$S = \sum_{i=1}^N (z_i - a + be^{-kt_i})^2. \tag{3.1}$$

The value of k obtained by this method of estimation will not in general be the same as that computable from $\hat{m}$ in (2.2) and used for the significance testing.

(2) Estimate k by means of (2.2) and the relation $m = e^{-k\Delta}$, then substitute this estimate into S of (3.1) and estimate a and b by means of least squares.

(3) Estimate k as in (2) and choose, as an estimate of a, the intercept of the "line of best fit" for 0_n. Then substitute these estimates of a and k into (3.1) and estimate b by means of least squares. In this case the estimate of b comes out to be:

$$\hat{b} = \sum_1^N e^{-\hat{k}t_i}(\hat{a} - z_i) \Big/ \sum_1^N e^{-2\hat{k}t_i} \tag{3.2}$$

where $\hat{a}$ and $\hat{k}$ are the estimates of a and k.

If the values, $t_1, t_2, \cdots, t_N$ are such that $t_{i+1} - t_i = \Delta$, $(i = 1, 2, \cdots, N-1)$, the following estimation procedure might be used.

(4) Let

$$y_j = z_{i+1} \qquad (i = 1, 2, \cdots, N-1),$$
$$x_j = z_i$$

and treat the $(N-1)$ pairs of values $(x_1, y_1; \cdots; x_{N-1}, y_{N-1})$ as a sample of size $(N-1)$. Using this sample, estimate k, a, and b in a manner similar to that in (2) or (3). It should be noted that this sample is not a random sample owing to the dependence among the $(N-1)$ elements.

The procedure in alternative (1) is very laborious and time-consuming. The procedure in (2) and (3) can be carried out quickly and easily. In (1) the method of least squares yields the same results as would be obtained from application of the method of maximum likelihood. Examples of estimation by procedures (3) and (4) are given in the next section.

4. Example. The accompanying table lists experimentally observed values of a property of a latex obtained at biweekly intervals. Using the first, third, *etc.*, quantities as x_j and the remaining ones as y_j, the sums of squares and products of deviations are found to be:

$$S_{xx} = .035510 \qquad \bar{x} = 0.9195$$
$$S_{xy} = .025645$$
$$S_{yy} = .023414 \qquad \bar{y} = .9365.$$

Substituting these values in equation (2.2) and computing the other constants from equation (2.1) we get: $m = 0.791596$, $a = 1.0009$, and $k = 0.1168$. The F' ratio is (2.3) 17.03. Entering Table I of [2], we find that for eight point pairs a value of $F' = 16.5$ may be expected only one time in one hundred. On excluding the possibility of negative values of m, this corresponds to the 0.5% significance level. The exponential relationship is thus concluded to be highly significant.

Evaluation of b by equation (3.2), method 3, gives 0.2560, if all 16 values are used. The equation calculated from the data is thus:

$$z = 1.0009 - 0.2560\, e^{-0.1163t}. \tag{4.1}$$

The alternative procedure, method 4, would be to use all the z_i points for the estimation of a and k. This leads to the following values of the computation quantities:

$$S_{xx} = \sum_{i=1}^{16} x_i^2 - x_{16}^2 = 0.052374; \qquad \bar{x} = 0.9223$$

$$S_{xy} = \sum_{i=1}^{15} x_i x_{i+1} = .036924$$

$$S_{yy} = \sum_{i=1}^{16} x_i^2 - x_1^2 = .035436; \qquad \bar{y} = .9381.$$

Note that the difference $S_{yy} - S_{xx}$ used in the formula for m cancels out all intervening squares between the first and last.

$$S_{yy} - S_{xx} = x_1^2 - x_{16}^2.$$

TABLE I

t weeks	z_i	t weeks	z_i	t weeks	z_i	t weeks	z_i
1	.776	9	.939	17	.942	25	.955
3	.852	11	.904	19	.938	27	.993
5	.850	13	.930	21	.979	29	.985
7	.869	15	.948	23	.975	31	1.013

However, the data excluded thereby are in effect included in the new S_{xy}.

The final values obtained by the fourth procedure are: $m = 0.796596$, $a = 1.0000$, and $k = 0.1137$. The writer does not know whether the peculiar transference of data from $S_{yy} - S_{xx}$ to S_{xy} characteristic of procedure 4 improves the accuracy of the fit or hurts it. It is his personal preference to use procedure 3.

5. Acknowledgement. The writer wishes to acknowledge with thanks his gratitude to Drs. T. W. Anderson, Jr. and David F. Votaw, Jr. for many suggestions and discussions concerning this problem and for much help in clarifying the presentation of the concepts.

REFERENCES

[1] CHARLES H. KUMMELL, *The Analyst* (Des Moines), Vol. 6 (1879), pp. 97–105.

[2] D. S. VILLARS AND T. W. ANDERSON, JR., "Some significance tests for normal bivariate distributions," *Annals of Math. Stat.*, Vol. 14 (1943), pp. 141–148.

ON THE POWER EFFICIENCY OF A t-TEST FORMED BY PAIRING SAMPLE VALUES

By John E. Walsh

Princeton University

1. Introduction. Consider two equal sized samples, one from a normal population with mean μ and the other from a normal population with mean ν. Let $x_1, \cdots, x_n$ be the sample values from the population with mean μ and $y_1, \cdots, y_n$ the values from the population with mean ν. If the two populations have the same variance and the two samples are independent, the most powerful tests for comparing μ and ν using these samples (one-sided and symmetrical two-sided) are based on the statistic

$$t_2 = \frac{[\bar{x} - \bar{y} - (\mu - \nu)]\sqrt{n(n-1)}}{\sqrt{\sum_1^n (x_i - \bar{x})^2 + \sum_1^n (y_i - \bar{y})^2}},$$

which has a Student t-distribution with $2n - 2$ degrees of freedom. Tests based on t_2 also have the desirable property of being invariant under permutation of the data in each sample.

Sometimes, however, it is useful to combine the sample values in the form

$$z_i = (x_i - y_i), \qquad (i = 1, \cdots, n).$$

Examples:

(a). When the samples are independent but it is not known that the two populations have the same variance (Behrens-Fisher problem).

(b). When there may be correlation between x_i and y_i, $(i = 1, \cdots, n)$, this correlation being the same for each value of i (i.e. x_i is independent of y_j if $i \neq j$ while each pair x_i, y_i, $(i = 1, \cdots, n)$, has the same normal bivariate distribution).

In both (a) and (b) the z_i are independently normally distributed with the same variance and mean $\mu - \nu$.

The Student t-test for comparing μ and ν using the z_i is based on the statistic

$$t_1 = \frac{[\bar{z} - (\mu - \nu)]\sqrt{n(n-1)}}{\sqrt{\sum_1^n (z_i - \bar{z})^2}} = \frac{[\bar{x} - \bar{y} - (\mu - \nu)]\sqrt{n(n-1)}}{\sqrt{\sum_1^n [x_i - y_i - (\bar{x} - \bar{y})]^2}},$$

which has a Student t-distribution with $n - 1$ degrees of freedom. These tests are not invariant under permutation of the data in each sample.

If it is true that all the sample values are independently distributed with the same variance σ^2, efficiency will be lost by using the test based on t_1 instead of the most powerful test based on t_2. The purpose of this note is to determine the power efficiency of the tests based on t_1 as compared with the corresponding tests based on t_2 for this case.

TABLE I

Power Function Values for the t_1 and t_2 Tests

Test	n	Approx. Efficiency	α	Approx. Values of Power Function			
				$\delta = \frac{1}{2}$	$\delta = 1$	$\delta = 1\frac{1}{2}$	$\delta = 2$
t_1	6	87%	.05	.276	.674	.933	.994
t_2	5.2		.05	.275	.672	.932	.994
t_1	6	82.5%	.025	.159	.486	.822	.970
t_2	4.95		.025	.160	.488	.823	.970
t_1	8	90%	.05	.355	.812	.985	
t_2	7.2		.05	.354	.813	.985	
t_1	8	86.5%	.025	.226	.674	.952	.998
t_2	6.9		.025	.225	.675	.951	.998
t_1	8	82%	.01	.112	.458	.843	.983
t_2	6.55		.01	.112	.457	.842	.983
t_1	10	92%	.05	.425	.898	.997	
t_2	9.2		.05	.425	.897	.997	
t_1	10	90%	.025	.289	.802	.988	
t_2	9		.025	.290	.803	.988	
t_1	10	85.5%	.01	.159	.626	.950	.999
t_2	8.55		.01	.159	.627	.950	.999
t_1	15	95.5%	.05	.579	.980		
t_2	14.3		.05	.579	.980		
t_1	15	93%	.025	.437	.950	1.000	
t_2	13.95		.025	.437	.949	1.000	
t_1	15	90%	.01	.278	.876	.998	
t_2	13.5		.01	.278	.876	.998	
t_1	25	98%	.05	.784	.999		
t_2	24.5		.05	.784	.999		
t_1	25	96%	.025	.670	.998		
t_2	24		.025	.670	.998		
t_1	25	94.5%	.01	.514	.992		
t_2	23.7		.01	.514	.992		

Consideration is limited to one-sided tests, which is not a serious limitation since any two-sided test can be considered as a combination of two one-sided tests. Table II contains approximate power efficiencies of one-sided tests for $n \geq 4$ at the significance levels $\alpha = .05, .025, .01$.

It is found that the efficiency of the t_1 test increases with the sample size but is high even for small size samples.

2. Outline of computations. The method of obtaining power efficiencies used here will be that outlined in [1]. Essentially this consists in computing the power function for the test based on t_1 and then adjusting the sample size for the corresponding test based on t_2 until its power function is approximately the same as for the t_1 test. The ratio of the sample size (perhaps fractional) of the adjusted t_2 test to that of the t_1 test is called the power efficiency of the t_1 test. Intuitively this efficiency measures the fraction of the total available information which is being used when the t_1 test is applied (since the t_2 test is most powerful).

TABLE II

Approximate Power Efficiencies for Given n and α

α \ n	4	5	6	7	8	9	10	15	25	∞
.05	82.5%	85%	87%	88.5%	90%	91%	92%	95.5%	98%	100%
.025	77%*	80%*	82.5%	84.5%	86.5%	88.5%	90%	93%	96%	100%
.01	73%	75.5%	78%	80%	82%	84%	85.5%	90%	94.5%	100%

* These values were obtained by comparison with the corresponding values for $\alpha = .05$ and .01.

It is easily seen from symmetry that a one-sided t_1 test of $\mu < \nu$ has the same power efficiency as the corresponding one-sided t_1 test of $\mu > \nu$. Thus it is sufficient to consider the one-sided tests of $\mu > \nu$.

The power function is found as a function of the parameter δ, where

$$\delta = \frac{\mu - \nu}{\sigma\sqrt{2}}.$$

Most of the approximate power efficiencies were determined by using the normal approximation given in [2] to compute the power function values. This approximation was used for fractional values of n. Table I contains the results of these computations for one-sided tests of $\mu > \nu$.

Exact values of the power function for integral values of n and $\alpha = .05, .01$ can be found from the tables in [3]. A comparison of the power function values obtained from the normal approximation with these exact values shows that, for $n \leq 6$, $\alpha = .01$ and $n \leq 4$, $\alpha = .05, .025$, the approximation underestimates the true values for small δ and overestimates for large δ. Although this combination of underestimation and overestimation tends to cancel out in the determina-

tion of power efficiencies, so that little error in power efficiencies would be expected if the approximation were used for $n = 6$, $\alpha = .01$ or $n = 4$, $\alpha = .05$, the efficiencies given in Table II for $n = 4$, $\alpha = .05$ and $n = 4, 6$, $\alpha = .01$ were obtained from the exact values by graphical interpolation and cross-interpolation.

Power efficiencies were not considered for $n < 4$ because of the difficulties of interpolation and the inexactness of the normal approximation in this range.

For $n = \infty$, t_1 and t_2 both have a normal distribution with zero mean and unit variance. Thus the power efficiency is 100% at all significance levels for this case.

These computations furnish approximate power efficiencies for $n = 6, 8, 10, 15, 25, \infty$ at $\alpha = .05, .025, .01$, and for $n = 4$ at $\alpha = .05$ and .01. The other approximate power efficiencies listed in Table II were obtained by graphical interpolation from these values.

The results of this note can be roughly summarized for $n \leq 15$ by stating that of the $2n$ sample values

(i). approximately 1.6 values are lost at the 5% significance level,

(ii). approximately 2.1 values are lost at the 2.5% significance level,

(iii). approximately 2.8 values are lost at the 1% significance level, if the tests based on t_1 are used instead of the corresponding tests based on t_2. Examination of Table I shows that the number of sample values lost decreases as n increases for $n > 15$.

REFERENCES

[1] JOHN E. WALSH, "On the power function of the sign test for slippage of means", *Annals of Math. Stat.*, Vol. 17 (1946), pp. 360, 361.

[2] N. L. JOHNSON AND B. L. WELCH, "Applications of the non-central t-distribution", *Biometrika*, Vol. 31 (1940), p. 376.

[3] J. NEYMAN, "Statistical problems in agricultural experimentation", *Roy. Stat. Soc. Suppl.*, Vol. 2 (1935), pp. 131, 132.

NOTE ON THE LIAPOUNOFF INEQUALITY FOR ABSOLUTE MOMENTS

BY MAURICE H. BELZ

The University of Melbourne

For a variate x measured from the mean of the population, the absolute moment of order r is defined by

$$\nu_r = \int_{-\infty}^{\infty} |x|^r \, dF(x),$$

where $F(x)$ is the cumulative distribution function. Treating r as continuous, we have

$$\frac{d\nu_r}{dr} = \int_{-\infty}^{\infty} |x|^r \log_e |x| \, dF(x),$$

the integral on the right existing if ν_{r+1} exists.

Write $y = \log_e \nu_r$. Then we have

$$\nu_r \frac{dy}{dr} = \int_{-\infty}^{\infty} |x|^r \log_e |x| \, dF(x),$$

$$\nu_r^2 \frac{d^2y}{dr^2} = \int_{-\infty}^{\infty} |x|^r \, dF(x) \cdot \int_{-\infty}^{\infty} |x|^r \log_e^2 |x| \, dF(x) - \left\{ \int_{-\infty}^{\infty} |x|^r \log_e |x| \, dF(x) \right\}^2$$

$\geqq 0$, by Schwarz's inequality.

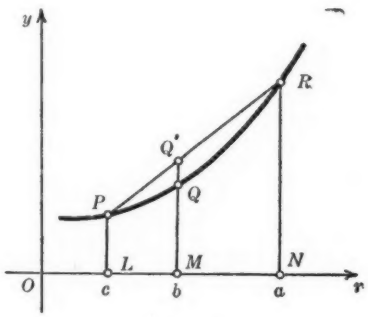


FIG. 1

It follows that the function y is convex (or exceptionally a straight line), and, on referring to the figure, it appears that

$$MQ \leqq MQ' \tag{1}$$

for all chords PR. If the abscissae of the points L, M, N are c, b, a, respectively, where $c \leqq b \leqq a$, the inequality (1) leads at once to the relation

$$\log_e \nu_b \leqq \frac{a-b}{a-c} \log_e \nu_c + \frac{b-c}{a-c} \log_e \nu_a .$$

Hence

$$\nu_b^{a-c} \leqq \nu_c^{a-b} \nu_a^{b-c},$$

which is the usual form of the Liapounoff Inequality.

REMARK ON THE NOTE "A GENERALIZATION OF WARING'S FORMULA"

BY T. N. E. GREVILLE

U. S. Public Health Service

Before submitting for publication the note "A generalization of Waring's formula," *Annals of Math. Stat.*, Vol. 15 (1944), pp. 218-219 the author made a diligent effort to ascertain, through correspondence with mathematicians and

actuaries both in this country and abroad, whether the generalized formula in question had been previously published, and none of the authorities communicated with knew of its prior publication. However, it has now come to his attention that the formula was published in essentially the same form by Hermite in the article "Sur la formule d'interpolation de Lagrange", *Journal für die Reine und Angewandte Mathematik* ("Crelle's Journal"), Vol. 84 (1878), pp. 70–79.

ABSTRACTS OF PAPERS

Presented Sept. 2–4, 1947, at the Yale meeting of the Institute

1. Estimation of Parameters in Truncated Pearson Frequency Distributions. A. C. COHEN, University of Georgia.

Given a truncated univariate Pearson frequency distribution, parameters of the complete distribution are required. Karl Pearson and Alice Lee, (*Biometrika*, Vol. 6 (1915), pp. 59–69) and R. A. Fisher, (*Introduction to Mathematical Tables*, Vol. 1, British Assn. Adv. Sci., 1931, pp. xxvi–xxxv), obtained solutions of the truncated normal distribution with a single *tail* missing. The present paper presents three general methods of solution applicable to any of the Pearson distributions. The first utilizes moments of a higher order than are required to characterize corresponding complete distributions. The order of the highest moment required is increased by one for each missing *tail*. The second method, applicable when only a single *tail* is missing, utilizes the terminal ordinate at the point of truncation and moments of the same order as required to characterize the complete distribution. The terminal ordinate is evaluated by successive approximations. The third method utilizes only the first two moments, but requires that the given distribution be further truncated and that moments be computed both before and after the additional truncations. This latter method can also be applied to complete distributions to avoid direct computation of third and fourth order moments.

2. Distribution of a Root of Determinantal Equation. D. N. NANDA, University of North Carolina.

The joint distribution of the roots of a determinantal equation was given by P. L. Hsu in 1939 and the distribution of any one of the roots was studied by S. N. Roy. The present paper, however, gives a different method of working out the distribution of any root, specified by its place in a monotonic arrangement. This method enables us to express the distribution of a root of a certain determinantal equation in terms of a linear combination of products of incomplete beta integrals and in terms of the distribution of a root of lower-order determinantal equations.

3. The Power of Certain Non-Parametric Tests of Independence. WASSILY HOEFFDING, University of North Carolina.

Several tests of independence have been proposed which are based on statistics depending only on the ranks of the sample values. Under the hypothesis H_0 of independence the distribution of such statistics does not depend on the form of the parent distribution. Two of these statistics, Spearman's rank correlation coefficient and Lindeberg-Kendall's statistic based on the number of inversions in the permutation of the ranks, are shown to be asymptotically normally distributed in samples from any population (the limiting normal distribution being singular in certain degenerate cases). The asymptotic distribution of these coefficients reveals that the corresponding tests of independence are inconsistent (in the sense that the probability of rejecting H_0 does not necessarily tend to 1 if H_0 is not true), and at least one of them is biased in the limit. It can be shown that at least for some sample sizes and some sizes of the critical region there do not exist unbiased tests of independence based on ranks. But there do exist rank tests of independence which are consistent, and hence unbiased in the limit. Examples of such tests are given.

4. Some Significance Tests for the Mean Using the Sample Range and Midrange. JOHN E. WALSH, Princeton University.

Consider a sample of size n, $(2 \leq n \leq 10)$, drawn from a normal population with mean μ. Let x_n be the largest value and x_1 the smallest value of the sample. Significance tests are developed to compare μ with a given hypothetical value μ_0 by use of the sample. These significance tests are based on the quantity $D = [\frac{1}{2}(x_1 + x_n) - \mu_0]/(x_n - x_1) = [(\text{sample midrange}) - (\text{hypothetical mean})]/(\text{sample range})$. One-sided and symmetrical tests are considered. Values of D_α such that $Pr(D > D_\alpha \mid \mu = \mu_0) = \alpha$ are computed for $\alpha = .05, .025, .01, .005$. These values of D_α can be used to obtain one-sided tests at the .05, .025, .01, .005 significance levels and symmetrical tests at the .10, .05, .02, .01 significance levels. Efficiencies are computed for one-sided tests at the .05 and .01 significance levels. The efficiency is at least 90% for $n \leq 6$ at the .05 significance level and for $n \leq 8$ at the .01 level. The range-midrange test can be applied without computation through the use of an easily constructed graph. The application of a test requires only the plotting of the sample point (x_1, x_n) on this graph.

5. Testing Compound Symmetry in a Normal Multivariate Distribution. DAVID F. VOTAW, JR., Princeton University.

Let $F(X)$ be the d.f. of a t-order vector variate $X (t \geq 3)$. Suppose the components of X are divided into mutually exclusive and exhaustive subsets. $F(X)$ is said to be *compound symmetric*, for the given division of its variates into subsets, if it is invariant over all permutations of its variates within these subsets. $F(X)$ is *completely symmetric* if the invariance holds over all permutations of its variates. If $F(X)$ is normal and compound symmetric, then within each subset of variates the means are equal, the variances are equal and the covariances are equal, and between any two subsets of variates the covariances are equal. Testing hypotheses of compound or complete symmetry in a normal $F(X)$ is of interest, for example, in studying psychological examinations and in medical research.

In this paper likelihood ratio criteria are developed for testing various hypotheses involving compound symmetry in regard to a normal distribution and to k normal distributions $(k \geq 2)$. Given that the corresponding null hypothesis is true, the moments of each criterion are obtained explicitly and the distribution of each criterion is identified as the product of independent beta variates (in the case of a single normal distribution, the distributions are given explicitly for $t = 3, 4$, and 5 for certain divisions of the variates into subsets). In a previous paper Wilks has given results on a very thorough study of the sampling theory of likelihood ratio criteria for various hypotheses involving complete symmetry in regard to a normal distribution.

6. Effects of Non-Normality at High Significance Levels. HAROLD HOTELLING, University of North Carolina.

The effects of non-normality in the underlying population on the probabilities of *significance* by customary statistical tests are not well understood, in spite of numerous attacks, both mathematical and experimental, on the problem. Chung's recent proof that the distribution of the Student ratio t has in samples from an arbitrary population a distribution approaching normality for large samples tends to confirm the common idea that non-normality makes little difference if only the sample is fairly large, but this holds only for a fixed range of values of t while the sample number N increases. The tail areas beyond a deviation which increases with N in certain ways often behave quite differently than in sampling from a normal population. If p is the probability that $| t | > t_0$ in samples of N from a normal population and p' is the corresponding probability for another population, it is shown that $\lim_{N \to \infty} \left\{ \lim_{t_0 \to \infty} (p'/p) \right\}$ may be zero or infinite or may take any finite value, even when the non-normal distribution involved is of simple and realistic continuous forms. The conditions that this limit be unity are concerned only with the *shoulders* of the population histogram, and have nothing to do with its moments or its

behavior at infinity or at its mean. This suggests that caution should be used in applying familiar tests with high significance levels; that further calculations should be directed toward making this caution quantitatively definite; and that the use of sample moments or cumulants cannot lead to the most appropriate criterion of non-normality for this purpose.

7. On the Problem of Similar Regions. E. L. LEHMANN, University of California, Berkeley, and HENRY SCHEFFÉ, University of California, Los Angeles.

If $X = (X_1, \cdots, X_n)$ is a set of random variables with a joint probability density depending on a set of parameters $\theta = (\theta_1, \cdots, \theta_m)$, and if $T = (T_1, \cdots, T_m)$ is a set of sufficient statistics for θ, then Neyman (*Phil. Trans. Roy. Soc. London*, Vol. 236 (1937), pp. 333–380) has proved that a region w in the space of X is similar with respect to θ if it has the following structure: The intersections $w(t)$ of w with the surfaces $T = t$ have the property that the conditional probability of the sample point X falling into w given that $T = t$ does not depend on t. In the present paper a necessary and sufficient condition is found for the regions with the above structure to be the only similar regions. This condition is shown to be satisfied for a certain class K of probability densities which contains as special cases all densities for which the totality of similar regions has been previously determined. In particular the partial differential equations which Neyman (*Annals of Math. Stat.*, Vol. 12 (1941), pp. 46–76) assumed were satisfied in his solution of the problem of similar regions are solved and it is shown that any density satisfying these equations belongs to the above class K.

8. Fourth Degree Exponential Function. L. A. AROIAN and MARGUERITE DARKOW, Hunter College.

It is shown that the fourth degree exponential function is supported by the Bernoulli probability function and the hypergeometric probability function as well as being the function for which the method of moments is the *best* method according to the criterion of maximum likelihood. In the general situation six moments, at most, are needed. The function is classified into two general groups depending on symmetry or asymmetry and each case is divided again into unimodal and bimodal distributions. Examples show that the function is very successful in graduating the main Pearson types and the Gram-Charlier Type A frequency function. Various generalizations of the exponential function are indicated. In addition to its wide generality, the greatest practical advantage of the new system is the simplicity of the numerical calculations.

9. A General Weak Limit Theorem for Independent Distributions. P. L. HSU, University of North Carolina. (Read by title.)

For every positive integer n let there be n distribution functions $F_{n1}(x)$, $F_{n2}(x)$, $\cdots$, $F_{nn}(x)$. Assume that $\lim_{n\to\infty} \mathrm{Max}_{1\leq j\leq n}\{1 - F_{nj}(x) + F_{nj}(-x)\} = 0$. Let $F(x)$ be the convolution $F_{n1}(x)*F_{n2}(x)* \cdots *F_{nn}(x)$. Let $\psi(t) = mit + \int_{-\infty}^{+\infty} [e^{itx} - 1 - itx/(1+x^2)](1+x^2)/x^2\, dG(x)$, with $G(x)\uparrow$ and $G(\infty) - G(-\infty) < \infty$. Let $F(x)$ be the (infinitely divisible) distribution law having $\exp \psi(t)$ as its characteristic function. In order to have $\lim_{n\to\infty} F_n(x) = F(x)$ at every continuity point of $F(x)$, it is necessary and sufficient that the following relations hold at every $x > 0$ such that $\pm x$ are continuity points of $G(y)$:

$$\text{(I)}\quad \lim_{n\to\infty} \sum_{j=1}^{n} \int_{|y|>x} dF_{nj}(y) = \int_{|y|>x} ((1+y^2)/y^2)\, dG(y),$$

$$\text{(II)}\quad \lim_{n\to\infty} \sum_{j=1}^{n} \left\{ \int_{|y|>x} y^2\, dF_{nj}(y) - \left(\int_{|y|<x} y\, dF_{nj}(y) \right)^2 \right\} = \int_{|y|<x} (1+y^2)\, dG(y),$$

$$\text{(III)}\quad \lim_{n\to\infty} \sum_{j=1}^{n} \int_{|y|<x} y\, dF_{nj}(y) = m + \int_{|y|<x} y\, dG(y) - \int_{|y|<x} (1/y)\, dG(y).$$

10. On the Maximum Partial Sums of Sequences of Independent Random Variables. K. L. CHUNG, Princeton University.

The asymptotic behavior of the maximum partial sums of a sequence of independent random variables is studied in this paper. Two groups of new limit theorems are established under general conditions. The first group deals with theorems of the *weak* type. The limiting distribution of the maximum partial sums is obtained with an estimate of the remainder, thus improving a recent result of Erdös and Kac. Another estimate is obtained for a different domain of variation, which plays an essential role in the sequel. These results correspond to the sharper forms of the central limit theorem. In the second group, theorems of the *strong* type are obtained, giving precise lower bounds (in the sense of probability) for the maximum partial sums. These results form the exact counterpart to the general form of the law of the iterated logarithm, due to Feller, which give the precise upper bounds. A summary of the main results and methods has appeared in *Proc. Nat. Acad. of Sci.*, Vol. 33 (1947), pp. 132–136.

11. Some Results on the Distribution of Quadratic Forms From Gaussian Stochastic Processes. (Preliminary report). HERMAN RUBIN, Cowles Commission.

If one considers the estimation of the parameters of a Gaussian stochastic process whose elements are continuous functions from the functional values over a finite interval, one often finds that certain parameters can be estimated exactly, and certain parameters can not. This result often depends on the distribution of quadratic functionals whose arguments are elements of the stochastic process under consideration. In this paper, it is shown that the elements of a certain class of quadratic functionals have distributions concentrated at a point, and that the elements of a different class do not; in this latter case, the characteristic function is computed.

12. Some Significance Tests for the Median which are Valid under Very General Conditions. (Preliminary Report) JOHN E. WALSH, Princeton University. (Read by title.)

Consider n independent values drawn from populations necessarily satisfying only: 1) Each population has a unique median. 2) The median has the same value φ for each population. 3) Each population is symmetrical. 4) Each population is continuous. (It is to be emphasized that no two of the values are necessarily drawn from the same population.) Significance tests are derived for φ on the basis of 1)–4). These significance tests are based on order statistics of certain combinations of order statistics, each combination being either a single order statistic of the n values or one-half the sum of two order statistics. The tests are invariant under permutation of the n values and reasonably efficient if the values represent a sample from a normal population. The significance levels are of the form $r/2^n$, $(r = 1, \cdots, 2^n - 1)$. Each value of r can be obtained for some one-sided significance test. Thus any significance level can be closely approximated if n is large. The major disadvantage of these tests is the limited number of suitable significance levels available for small values of n. This disadvantage is partially eliminated by the development of tests which have a specified significance level if the values are a sample from a normal

population and a significance level bounded near this specified value if only 1)–4) necessarily hold. Results based on 1)–4) are applied to several well known statistical problems: Tests are obtained for the mean on the basis of a large number of independent values from populations having the mean but little else in common. Also generalized results are obtained for the Behrens-Fisher problem, quality control, slippage tests, the sign test and cases where some of the n values are dependent.

13. Loss of Information in t-tests with Unbalanced Samples. (Preliminary Report) JOHN E. WALSH, Princeton University. (Read by title.)

Consider two normal populations, the first with mean a_1 and variance σ_1^2, the second with mean a_2 and variance σ_2^2, while σ_1/σ_2 has a known value C. If the hypothesis $a_1 = a_2$ is to be tested by a t-test (one-sided or symmetrical) using n_1 sample values from the first population and n_2 values from the second population ($n_1 + n_2 = n$, fixed), it is shown that this experiment is most powerful when $n_1/n_2 = \sigma_1/\sigma_2$ (integer considerations neglected). The t-tests satisfying this condition will be referred to as balanced t-tests. Thus information will be lost by not using a balanced experiment. A quantitative measure of the information lost by using given values of n_1 and n_2 is determined by the total sample size m, ($m_1 + m_2 = m$), of the balanced t-test (same significance level) which has approximately the same power. Then $n - m$ sample values are wasted by using (n_1, n_2) rather than (m_1, m_2), i.e. only $100m/n\%$ of the information obtainable per observation is used by (n_1, n_2). A symmetrical t-test with significance level 2α has the same value of m as a one-sided t-test with significance level α. For one-sided t-tests with significance level α: $m \doteq \frac{1}{2}(B + \sqrt{B^2 - 8A})$, where $B = 2 + A + K^2/2$, $A = (C+1)^2[1 - K^2/2(n-2)][C^2/n_1 + 1/n_2]^{-1}$, and K_α is the standardized normal deviate exceeded with probability α. This approximation to m is valid for $m \geq 5$ if $\alpha = .05$, $m \geq 6$ if $\alpha = .025$, $m \geq 7$ if $\alpha = .01$, $m \geq 8$ if $\alpha = .005$. (A fractional value of m represents an interpolated measure of the sample size of the equivalent balanced experiment.)

14. Some Theorems on the Bernoullian Multiplicative Process. T. E. HARRIS Princeton University. (Read by title.)

A single entity may have j descendents with probability P_j, ($j = 0, 1, 2, \cdots$). Each *first generation* entity has then the same procreative probabilities, etc. Let

$$f(s) = p_0 + p_1 s + \cdots$$

If z_n is the number of entities in the nth generation, it is known that $P(z_n = j)$ is given by the coefficient of s^j in the nth iterate $f[f \cdots (f)] = f_n(s)$. Let $Ez_1 = x$, $1 < x < \infty$. Conditions are given insuring that as $n \to \infty$ the cumulative distribution of the variate z_n/x^n approaches a limit-function which is absolutely continuous except for a possible single jump. Let $g(u)$ be the corresponding frequency function. If $f(s)$ is a polynomial of degree k, let $q = \log_x k/(\log_x k - 1)$. Otherwise, $q = 1$. Then $g(u) \cdot \exp\{u^{q+\epsilon}\}$ [is, is not] summable $(0, \infty)$ according as ϵ is [negative, positive]. Behavior of $g(u)$ near $u = 0$ is also considered. Special cases are considered where $g(u) = \text{constant} \cdot u^{1/m-1} \cdot e^{-u/m}$, m a positive integer. Maximum likelihood estimates for the parameters $p_0, p_1, \cdots$, and x are obtained as functions of n successive values $z_1, z_2, \cdots, z_n$. Consistency, in a certain sense, is proved. A specialized method is given for finding the moment-generating function of the variate N, the smallest value of n such that $z_n = 0$.

NEWS AND NOTICES

Readers are invited to submit to the Secretary of the Institute news items of interest

Personal Items

Dr. George E. Albert has been appointed to an associate professorship at the University of Tennessee.

Dr. T. W. Anderson, Jr. has been promoted to an assistant professorship in the Department of Mathematical Statistics at Columbia University. He is on leave the first half of the 1947–48 academic year at the Institute of Actuarial Mathematics and Mathematical Statistics, Stockholm University as a Guggenheim Fellow. During the second half of the academic year he will be at Cambridge University.

Associate Professor Max Astrachan has been promoted to a full professorship at Antioch College, Yellow Springs, Ohio.

Associate Professor T. A. Bancroft, who has been at the University of Georgia, Athens, Georgia, is now with the Statistical Laboratory, Alabama Polytechnic Institute, Auburn, Alabama.

Dr. M. S. Bartlett of Cambridge University has been appointed as Professor of Mathematical Statistics at the University of Manchester, Manchester, England. The position is a newly created one. Professor Bartlett indicates that this position is believed to be the first official professorship in mathematical statistics in England.

Professor M. A. Brumbaugh has accepted a position with the Bristol Laboratories Inc., Syracuse 1, New York.

Dr. Donald A. Darling has been appointed Research Associate at Cornell University.

Professor D. B. DeLury of the Virginia Polytechnic Institute has accepted a position with the Ontario Research Foundation, 43 Queens Park, Toronto 5, Canada.

Professor Abel Gauthier of the University of Montreal has been appointed Head of the Institute of Mathematics and Assistant-Secretary of the Faculty of Science at that institution.

Dr. Casper Goffman, former assistant professor in the Mathematics Department, University of Kentucky, is now in the Mathematics Department, University of Oklahoma, Norman, Oklahoma.

Mr. Philip Hardy has returned to the General Electric Company at Warren, Ohio after serving at Wright Field.

Dr. Carl F. Kossack, who has been with the Navy Department in Washington, D. C. as an Air Intelligence Specialist, has accepted an associate professorship in the Department of Mathematics at Purdue University.

Mr. Frank Jones Massey, Jr. is now teaching in the Department of Mathematics, University of Maryland, College Park, Maryland.

Dr. William Burton Michael, who has been Lecturer in Mathematics, Psychology and Educational Psychology at the University of Southern California, has now accepted an assistant professorship in the Department of Psychology, Princeton University. He is also a member of the Research Department, College Entrance Examination Board at Princeton.

Mr. Bernard Ostle, a former teaching assistant, School of Business Administration, University of Minnesota, is now at Iowa State College, Ames, Iowa.

Mr. Maurice H. Quenouille, who was formerly with the Rothamsted Experimental Station, Harpendon, Herts, England, has accepted the position of Lecturer in Statistics, Marischal College, University of Aberdeen, Scotland.

Dr. James A. Rafferty left the Department of Pathology, University of Rochester in June and has been appointed Chief of the Department of Statistics, Air University, School of Aviation Medicine, Randolph Field, Texas.

Miss Mary Ann Savas has accepted a position with General Motors, Detroit, Michigan.

Professor George J. Stigler, formerly with Brown University, is now teaching in the Department of Economics, Columbia University, New York, New York.

Professor E. L. Welker has resigned an associate professorship in mathematics at the University of Illinois to become Associate in Mathematics in the Bureau of Medical Economic Research of the American Medical Association.

Mr. Sol M. Wezelman, who completed his master's degree in actuarial science at the University of Michigan in June, has accepted a position as Assistant Actuary in the North Dakota State Department of Insurance, Bismarck.

Dr. Bertram Yood has received his doctorate at Yale and is now on the staff at Cornell University.

Mr. Earl K. Yost, Jr. has accepted a position with the General Electric Co. at the Hanford Engineering Project, Richland, Washington.

Professor James G. Smith, of Princeton University, died at Princeton on November 28, 1946.

Beginning with the October issue, the quarterly journal *Mathematical Tables and Other Aids to Computation* will publish a new feature section, "Automatic Computing Machinery," designed to disseminate information and news on research and development in the field of high-speed automatic calculating machinery. Material should fall under the general headings of Bibliography, Technical Developments, Discussion (including correspondence), and News. Contributions to this section are invited and should be addressed to Dr. E. W. Cannon, Head of the Mathematics Group, Machine Development Laboratory, National Bureau of Standards, Washington, D. C.

Institute of Numerical Analysis Established

Plans have been completed for the establishment of one of the newest units of the National Bureau of Standards—the Institute of Numerical Analysis—at the

University of California at Los Angeles, according to an announcement by Dr. Edward U. Condon, Director of the Bureau.

One of the giant high-speed electronic computing machines, now under development by the Bureau of Standards, will be installed at the Institute when completed. Design specifications call for high memory capacity and automatically sequenced mathematical operations from start to finish at speeds attainable only with electronic equipment.

The Institute has two primary functions. The first is research in applied mathematics aimed at developing methods of analysis which will extend the use of the high-speed electronic computers. The second is to act as a service group for Western industries, research institutions, and government agencies. The service function will include not only the use of the machines for problem solving but also assistance in the formulation of problems in applied mathematics of the more complex and novel types. Service operations are to be initiated immediately, using the latest types of commercially available computing equipment.

The decision to locate the Institute at the University of California at Los Angeles was made after a nation-wide survey by the National Bureau of Standards. Centers in the East and Middle West were considered as well as the Far West, but Los Angeles, it was decided, offered the widest range of possibilities for an Institute of Numerical Analysis. Concentration of aircraft industries and the presence of several major scientific institutions were critical in the choice of Los Angeles.

Election of Fellows

The Board of Directors announced at the Yale Meeting the election of the following members as Fellows of the Institute: Theodore W. Anderson, Jr., Alexander C. Aitken, David H. Blackwell, Georges Darmois, Ragnar Frisch, Robert C. Geary, Frederick Mosteller, Gerhard Tintner, Charles P. Winsor and John Wishart.

New Members

The following persons have been elected to membership in the Institute (June 1 to August 29, 1947)

Baldwin, Helen Mildred, B.S. (Cornell) Research Associate in Statistics, Atomic Energy Project, *215 Avenue C, Rochester 5, N. Y.*

Blunk, Paul M., A.B. Teaching asst. and grad. student, Univ. of Calif., *Box 532, Fair Oaks, Calif.*

Bowden, George Edwin, B.S. (Duke) Teaching asst., Math. Dept., White Hall, Cornell Univ., Ithaca, N. Y.

Bradley, Ralph Allan, M.A. (Queen's Univ.) Grad. student, Univ. North Carolina, *Wellington, Ontario, Canada.*

Burton, Kenneth John, Head of Statistics Section, British Employers' Confederation, *16 Rutherwyke Close, Ewell, Surrey, England.*

Carlson, Phillip G., Jr., A.M. (Columbia) *148 Cornell Street, Roslindale 31, Mass.*

Carol, Bernard, M.S.E. (Columbia) Graduate student at Columbia Univ., *15 West 96th Street, N. Y.*

Clark, Sidney B., B.A. (George Wash. Univ.) Statistician, Public Roads Administration, *2728 Porter St., N.W., Wash., 8, D. C.*

Danielson, Theresa, M.A. (Univ. of Ill.) Mathematician at Brookhaven National Laboratory, *3512 Cambridge Ave., New York 63, N. Y.*

Dineen, Russell D. F., B.A. (Univ. of Delaware) Graduate student at Univ. of Delaware, *1318 French Street, Wilmington, Delaware.*

Diver, M. L., M.E. (Purdue) Consulting Engineer, *P.O. Box 1016, San Antonio 6, Texas.*

Erasmus, Josias C., M.S.E. (Univ. of Stellenbasch, South Africa) Research Officer, Grootfontein College of Agriculture, Middelburg, C-P, South Africa.

Gottlieb, Morris J., Ph.D. (Wash. Univ., St. Louis) Member of the Institute for Advanced Study, Washington University, St. Louis, Mo.

Greenwood, Joseph Arthur, A.B. (Harvard) Student at Harvard University, *66 Oxford St., Cambridge 38, Mass.*

Gysbers, Jack C., M.A. (Univ. of Calif.) Teaching asst., Dept. of Math., Univ. of Calif., *2029 Berkeley Way, Berkeley 4, Calif.*

Haskind, Mina, B.S. (Brooklyn College) Student at Brooklyn College, *763 Eastern Parkway, Brooklyn 13, New York.*

Hauser, Dr. Philip M., Ph.D. (Univ. of Chicago) University of Chicago, Chicago 37, Ill.

Hoyt, Cyril J., Ph.D. (Univ. of Minn.) Research Associate, Dept. of Education, University of Chicago, Chicago, Ill.

Kern, Enrique Roberto, First Assistant, Institute of Biometry, Univ. of Buenos Aires, Rivadavia 8854, Buenos Aires, Argentina.

Mark, Abraham M., Ph.D. (Cornell) Mathematics Department, Univ. of Wisconsin, Madison, Wisconsin.

Moss, George G., II, B.A. (St. John's College, Annapolis) Actuarial Statistician, Metropolitan Life, *2771 Morris Ave., N. Y. 58, N. Y.*

Phillips, Bernard E., A.M. (Columbia) *Box 147, Cathedral Station, New York 25, N. Y.*

Radvanyi, Laszlo, Ph.D. (Univ. of Heidelberg) Professor of Economics, National Univ. of Mexico, *Donato Guerra 1, desp. 207, Mexico, D. F.*

Richardson, John M., Ph.D. (Cornell) Member of Technical Staff, Bell Telephone Laboratories, Inc., Murray Hill, New Jersey.

Royston, Robert W., M.S. (Univ. of Mich.) Asst. Prof., Math. Dept., Wash. & Lee Univ., *117 W. Washington St., Lexington, Virginia.*

Savas, Mary A., A.B. (Univ. of Mich.) Student at Univ. of Mich., *524 E. Second St., Monroe, Mich.*

Shepard, David H., A.B. (Univ. of Mich.) Research Analyst, Army Security Agency, *505 Randolph Street, Falls Church, Virginia.*

Throdahl, Monte C., B.S. (Iowa State College) Research Chemist in Charge of Rubber Lab., Monsanto Chemical Co., Nitro, West Virginia.

Uchytil, Jan, Doctor of Science, Chief of Production Control Dep. in Central Federation of Czech. Industry, *Praha II, Prikopy 14, Czech.*

Vergara, Jose, Doctor of Engineering, (Madrid) Professor of Economics, Madrid, Chief of the Bureau of Statistics, Dept. of Agric., Madrid. *5262 S. Blackstone Ave., Chicago 37, Illinois.*

Wei, Dzung-shu, Ph.D. (Univ. of Iowa) Prof. and Head of Math. Dept., St. John's Univ., Shanghai, *129 East 10th St., New York 3, N. Y.*

Wolfson, Jacob, B.A. (New York College) Statistician, Social Security Adm., *845 Brunswick Road, Essex, Maryland.*

REPORT ON THE NEW HAVEN MEETING OF THE INSTITUTE

The Tenth Summer Meeting of the Institute of Mathematical Statistics was held at Yale University, New Haven, Connecticut, Tuesday, September 2 through Thursday, September 4, 1947. The meeting was held in conjunction with the summer meetings of the American Mathematical Society and the Mathematical Association of America. The following 150 members of the Institute attended the meeting:

C. B. Allendoerfer, R. L. Anderson, H. E. Arnold, L. A. Aroian, H. M. Bacon, J. L. Barnes, W. D. Baten, R. E. Bechhofer, A. A. Bennett, Joseph Berkson, D. H. Blackwell, C. I. Bliss, Colin Blyth, Jr., A. E. Brandt, G. M. Brown, R. H. Brown, O. P. Bruno, P. T. Bruyere, Mrs. P. T. Bruyere, J. H. Bushey, B. H. Camp, G. C. Campbell, Uttam Chand, K. L. Chung, W. G. Cochran, A. C. Cohen, Jr., E. P. Coleman, T. F. Cope, G. M. Cox, C. C. Craig, E. L. Crow, H. B. Curry, G. B. Dantzig, M. D. Darkow, B. B. Day, Bernard Dimsdale, C. E. Dieulefait, C. W. Dunnett, Churchill Eisenhart, L. R. Elveback, M. W. Eudey, H. P. Evans, William Feller, C. D. Ferris, M. M. Flood, R. M. Foster, H. A. Freeman, J. E. Freund, H. P. Geiringer, M. J. Gottlieb, J. Arthur Greenwood, Evelyn Groosman, F. E. Grubbs, H. T. Guard, P. R. Halmos, Max Halperin, M. H. Hansen, B. I. Hart, Mina Haskind, Wassily Hoeffding, R. H. Hoskins, Harold Hotelling, A. S. Householder, Jaroslav Janko, Irving Kaplansky, Leo Katz, Oscar Kempthorne, E. M. Kennedy, W. L. Kichline, C. J. Kirchen, L. F. Knudsen, H. S. Konijn, C. F. Kossack, Jack Laderman, H. G. Landau, E. L. Lehmann, R. A. Leibler, Walter Leighton, Jr., F. C. Leone, Joseph Lev, Howard Levene, Julius Leiblein, Arthur Linder, S. B. Littauer, E. D. Lowry, H. F. MacNeish, P. J. McCarthy, John Mandel, H. B. Mann, Sophie Marcuse, F. J. Massey, Margaret Merrell, E. B. Mode, M. E. Moore, Frederick Mosteller, D. N. Nanda, P. M. Neurath, M. C. Neurdenburg, G. E. Noether, M. L. Norden, H. W. Norton, P. S. Olmstead, A. L. O'Toole, E. R. Ott, T. E. Oxtoby, Edward Paulson, M. P. Peisakoff, G. B. Price, J. A. Rafferty, L. J. Reed, C. J. Rees, P. R. Rider, John Riordan, H. E. Robbins, Milton da Silva Rodrigues, A. C. Rosander, Ernest Rubin, Herman Rubin, Frank Saidel, M. M. Sandomire, Arthur Sard, Max Sasuly, F. E. Satterthwaite, E. D. Schell, Jack Sherman, Rosedith Sitgreaves, Andrew Sobczyk, Milton Sobel, Herbert Solomon, Mortimer Spiegelman, Arthur Stein, Henry Teicher, R. M. Thrall, Gerhard Tintner, M. N. Torrey, J. W. Tukey, D. F. Votaw, Jr., Abraham Wald, H. M. Walker, J. E. Walsh, R. M. Walter, J. H. Watkins, Dzung-shu Wei, E. S. Weiss, S. S. Wilks, C. P. Winsor, H. O. Wold, Jacob Wolfowitz, C. A. Wright, Bertram Yood.

The Tuesday afternoon session was devoted to a symposium on 2 x 2 tables with Professor Lowell J. Reed of Johns Hopkins University serving as chairman. Addresses were given on *Tests of Significance* by Dr. Churchill Eisenhart, National Bureau of Standards; *Estimation* by Dr. Charles P. Winsor, Johns Hopkins University and *Non-Standard Cases* by Dr. Joseph Berkson, Mayo Clinic. Discussants were Mr. William F. Taylor, Dr. Frederick Mosteller, Professor David H. Blackwell and Professor John W. Tukey. The attendance was approximately 130.

The first Wednesday morning session was devoted to contributed papers. Professor John W. Tukey of Princeton University presided. The attendance was approximately 85. The following three papers were presented:

1. *Estimation of Parameters in Truncated Pearson Frequency Distributions.* Professor A. C. Cohen, University of Georgia.

2. *Distribution of a Root of a Determinantal Equation.*
   Mr. D. N. Nanda, University of North Carolina.
3. *The Power of Certain Non-Parametric Tests of Independence.*
   Dr. Wassily Hoeffding, University of North Carolina.

The second Wednesday morning session was held with Professor Will Feller, President of the Institute, presiding. Professor R. A. Fisher, University of Cambridge, gave the address under the title *The Fitting of Gene Frequencies to Data for Genotypes.* The attendance was approximately 160.

The membership business meeting of the Institute was held at 9:15, Thursday morning, in 102 Chittenden Hall with President Feller presiding. The attendance was approximately 55. It was voted to make certain changes in the By-Laws and in particular to raise the due to $7.00 per year. (An exception is made for those living outside the Western Hemisphere.) Morris Hansen, Chairman of the Committee on Planning and Development, initiated a lively discussion with reference to desirable changes in the Constitution.

On Thursday morning at 10:30, with President Feller presiding, Professor A. Wald of Columbia University presented the Henry Lewis Rietz Lecture on *Sequential Estimation and Multi-Decisions.* The attendance was approximately 150.

A joint session with the American Mathematical Society was held early Thursday afternoon at which Professor S. S. Wilks of Princeton University gave a lecture on *Sampling Theory of Order Statistics.* Professor Harold Hotelling of the University of North Carolina was the presiding officer. The attendance was approximately 300.

This session was followed by another joint session with the American Mathematical Society which was devoted to contributed papers. Professor John W. Tukey presided at this session and the attendance was approximately 115. The following seven papers were presented:

1. *Some Significance Tests for the Mean Using the Sample Range and Midrange.*
   Mr. John Walsh, Princeton University.
2. *Testing Compound Symmetry in a Normal Multivariate Distribution.*
   Dr. David F. Votaw, Jr., Princeton University.
3. *Effects of Non-Normality at High Significance Levels.* Professor Harold Hotelling, University of North Carolina.
4. *On the Problem of Similar Regions.*
   Dr. Erich L. Lehmann, University of California, Berkeley and Professor Henry Scheffe, University of California at Los Angeles.
5. *The Fourth Degree Exponential Function.*
   Dr. Leo A. Aroian and Professor Marguerite Darkow, Hunter College.
6. *On the Maximum Partial Sums of Sequences of Independent Distributions.*
   Dr. K. L. Chung, Princeton University.
7. *Some Results on the Distribution of Quadratic Forms from Gaussian Stochastic Processes.*
   Mr. Herman Rubin, Cowles Commission.

The following four papers were presented by title:

8. *A General Weak Limit Theorem for Independent Distributions.*
   Professor P. L. Hsu, University of North Carolina.

9. *Some Significance Tests for the Median which are Valid under Very General Conditions* (Preliminary Report).
   Mr. John E. Walsh, Princeton University.
10. *Loss of Information in t-tests with Unbalanced Samples* (Preliminary Report).
    Mr. John E. Walsh, Princeton University.
11. *Some Theorems on the Bernoullian Multiplicative Process.*
    Mr. T. E. Harris, Princeton University.

Abstracts of all these papers appear elsewhere in this issue of the *Annals.*

A beer party in honor of the foreign statisticians attending the meeting was held in the dining room of Saybrook College on Wednesday evening. A joint dinner with the American Mathematical Society and the Mathematical Association of America was held on Thursday evening.

C. C. CRAIG,
Acting Secretary.

BIOMETRIKA

A Journal for the Statistical Study of Biological Problems

Vol. XXIV, Parts I, and II, January, 1947

CONTENTS

The subscription price, *payable in advance*, is 45s. inland, 54s. export (per volume including postage). Cheques should be drawn to Biometrika and sent to "The Secretary, Biometrika Office, Department of Statistics, University College, London, W.C. 1." All foreign cheques must be in sterling and drawn on a bank having a London agency.

ECONOMETRICA

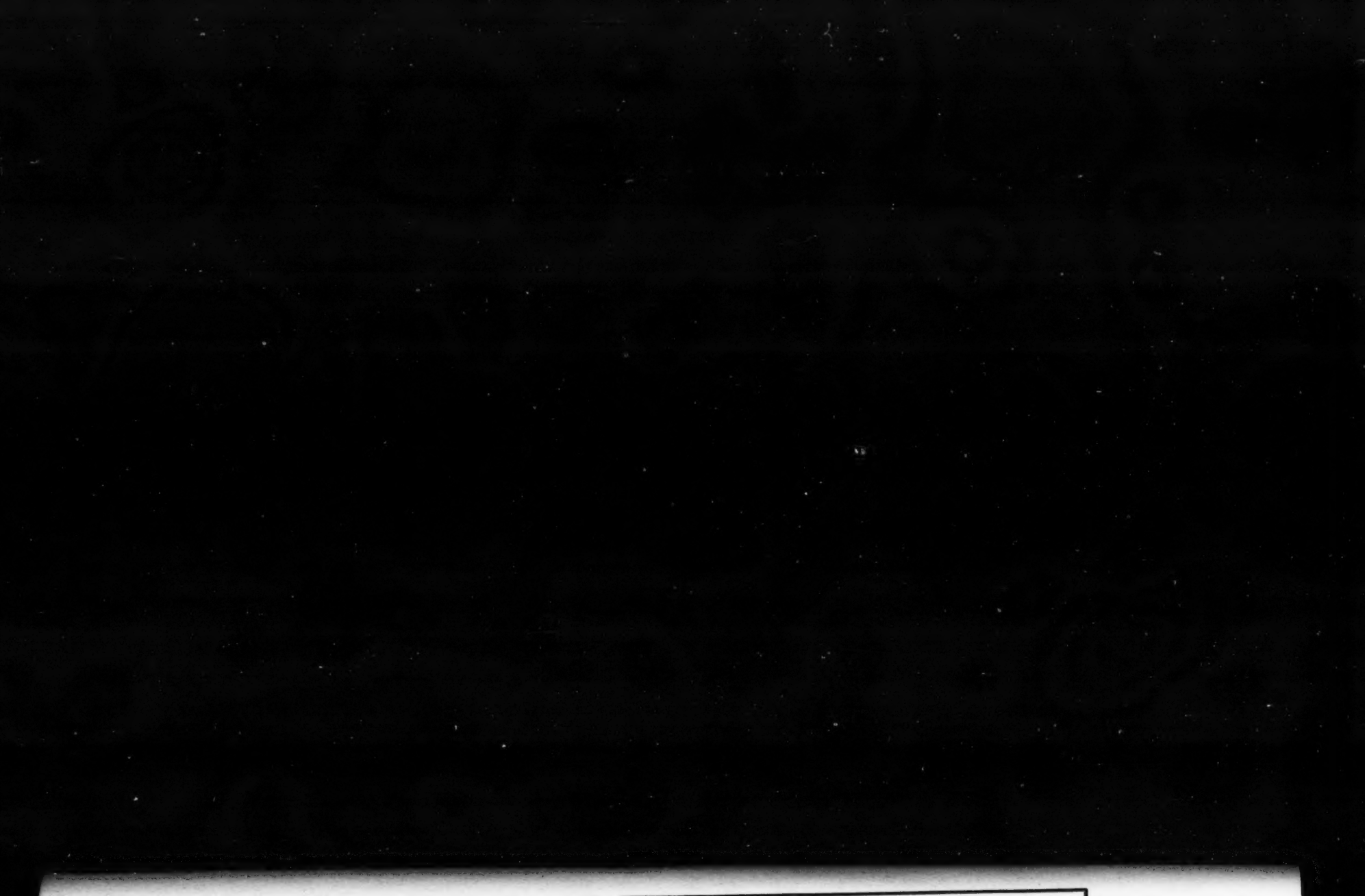
Journal of the Econometric Society

Contents of Vol. 15, No. 3, July, 1947

Published Quarterly Subscription: $7.00 yer year

The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.

Subscriptions to *Econometrica* and inquiries about the work of the Society and the procedure in applying for membership should be addressed to Alfred Cowles, Secretary and Treasurer, The Econometric Society, The University of Chicago, Chicago 37, Illinois.



THE INSTITUTE OF MATHEMATICAL STATISTICS

(Organized September 12, 1935)

OFFICERS FOR 1947

President:

WILL FELLER, Cornell University, Ithaca

Vice-Presidents:

J. H. CURTISS, National Bureau of Standards

M. H. HANSEN, Bureau of the Census

Secretary-Treasurer:

PAUL S. DWYER, University of Michigan, Ann Arbor

The purpose of the Institute of Mathematical Statistics is to stimulate research in the mathematical theory of statistics and to promote coöperation between the field of pure research and the fields of application.

Membership dues including subscription to the ANNALS OF MATHEMATICAL STATISTICS are $5.00 per year. The dues and inquiries regarding membership in the Institute should be sent to the Secretary-Treasurer of the Institute.